DICHOTOMIZING ILLNESS FROM CARDIOVASCULAR AND LOCOMOTOR ACTIVITY TIME SERIES

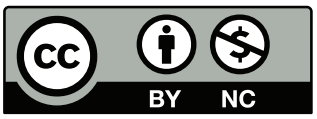
A Dissertation
Presented to
The Academic Faculty

By

Erik Reinertsen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Biomedical Engineering
at the Georgia Institute of Technology
and Emory University School of Medicine

August 2018



DICHOTOMIZING ILLNESS FROM CARDIOVASCULAR AND LOCOMOTOR ACTIVITY TIME SERIES

Approved by:

Gari D. Clifford, DPhil
Departments of Biomedical Informatics and Biomedical Engineering,
Emory University School of Medicine, and Georgia Institute of Technology

Shamim Nemati, PhD
Department of Biomedical Informatics,
Emory University School of Medicine

Eberhard Voit, PhD
Department of Biomedical Engineering,
Georgia Institute of Technology

Lee Cooper, PhD
Departments of Biomedical Informatics and Biomedical Engineering,
Emory University School of Medicine, and Georgia Institute of Technology

Amit J. Shah, MD, MSCR
Division of Cardiology, Department of Medicine,
Emory University School of Medicine
Department of Medicine,
Rollins School of Public Health at Emory University

Date Approved: 24th July, 2018

DEDICATION

To my parents and Chai Yoon.

ACKNOWLEDGEMENTS

This thesis was possible thanks to years of support, guidance, and inspiration from Prof. Gari Clifford. The Department of Biomedical Informatics has grown under his leadership, and is poised to contribute to how computation impacts research and clinical care across many specialties. The author also wishes to thank other mentors, advisors, and colleagues at Emory and Georgia Tech, including:

- Prof. Shamim Nemati: who showed me the importance of attention to detail, independent thinking, and first principles.
- Prof. Amit Shah: who provided contagious enthusiasm, enduring optimism, and valuable clinical perspectives.
- Prof. Eberhard Voit: who through attention to detail and great dedication to education helped develop me as a teacher and systems thinker.
- Prof. Lee Cooper: who was encouraging and always supportive.
- Supreeth Shashikumar and Ayse Cakmak: for conversations about work and life on nights and weekends, and resolving hundreds of code issues on GitHub. Your greatest work still lies ahead of you, and your careers will be inspirational.
- Barbara Birt: for her kind attitude and dedication to helping us all achieve more. Rest in peace.

The author would also like to acknowledge the following funding bodies, which provided project or fellowship funding for the research presented in this thesis:

- National Science Foundation Award 1636933
- National Institutes of Health Grant K23 HL127251
- National Institutes of Health Grant P50 HL117929

- National Institutes of Health Grant K01 ES025445
- National Institutes of Health Grant R01 HL136205

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

The author also thanks Drs. John M. Kane, Georgios Petrides, and Yashar Behzadi for the collection of the schizophrenia data, and Prof. Qiao Li for the pulse detection and signal quality code used in this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis addresses the issue of automated evaluation of severity of illness in psychiatric populations. In particular, given that both physiology and locomotor activity have been shown to be modified during mental illness, this work analyses the potential for the use of these measures to assess the discrimination of mental illness using supervised learning algorithms. In particular it examines the discriminatory power of information in heart rate time series and locomotor activity in three ways: 1) using multiple time scales (from minutes to several days), 2) during specific times (as a proxy for context) and 3) using interactions between locomotor and physiological time series.

This thesis is comprised of four parts: 1) a review of past work, 2) classification of mental illness using features from quiescent segments of HR, 3) classification of mental illness using features from both heart rate and locomotor activity time series over varying time scales, and 4) evaluation of coupling and interactions between heart rate and activity as features for classifying illness.

In Part 1), the body of work upon which this thesis builds is summarized in a review of digital sensors for neuropsychiatric illness. First, the two specific mental illnesses of focus are discussed: schizophrenia and PTSD. Heart rate variability (HRV) and locomotor activity, as well as relevant metrics and features therein are reviewed. The growing literature on digital sensors for monitoring neuropsychiatric illnesses is surveyed, with a focus on passive monitoring and analyses of HR and locomotor activity, feature extraction, and classification or regression of clinically relevant outcomes. In Part 2), features from heart rate time series data are used to train a classifier to distinguish post-traumatic stress disorder from controls subjects. This work explores the hypothesis that data from quiescent (low activity) segments will be more useful for discrimination than data from other segments during the 24-hour recording. This is driven by the knowledge that sleep minimizes exogenous sources of HRV,

such as social routine and physical activity. Dysautonomia detectable via alterations in HRV measures such as LF and HF power may thus be amplified during these quiescent segments. Classification is shown to be improved by segmenting the data using low heart rate segments as a proxy for the most restful period of sleep.

In Part 3), the work explores the hypothesis that information relevant to pathologically altered physiology and behavior varies with time scale. Features from both heart rate and locomotor activity data recorded over several days are used to train a classifier to distinguish subjects with schizophrenia from healthy controls. The time scale (e.g. window length) of data is varied and found to affect classifier performance, which has a direct relevance to the practical usage of the classifier.

In Part 4), the work explores the hypothesis that information between signals is altered in mental illness and relatively less altered in cardiovascular illness, and that this information is useful in a machine learning approach to discriminate patients from controls. Interactions between heart rate and locomotor activity are evaluated using information theoretical approaches, and found to contribute significantly to the classification of schizophrenia over combining univariate approaches, and differently to the classification of mental versus cardiovascular illness.

In summary, this thesis demonstrates that physiological data and locomotor activity data, over multiple time scales, independently provide discriminatory power in evaluating psychiatric conditions. Furthermore, the interaction between the two domains (movement influencing physiology and vice-versa) provides significant additional discriminatory power.

CHAPTER 1
INTRODUCTION

1.1 Motivation

Neuropsychiatric illness comprises 13-16% of the total global burden of disease measured in disability life-adjusted years (DALYs) for all ages, which exceeds the burden of cardiovascular disease or cancer (Vigo et al. 2016). One in four people in the world will be affected by mental or neurological disorders at some point in their lives, yet only a small fraction of the 450 million people affected will receive treatment due to pervasive underdiagnosis, a lack of trained healthcare professionals, stigma, and other reasons (Sayers 2001). These illnesses are more prevalent among older people and will contribute even more to overall global disease as life expectancy improves. The burden of mental and substance use disorders increased by 37% between 1990 and 2010, which for most disorders was driven by population growth and aging (Whiteford et al. 2013). The prevalence of dementia continues to rise, and by 2050 an estimated 13.8 million Americans will have Alzheimer's disease (AD; see A1 for definitions of abbreviations and acronyms used in this thesis) or another dementia. In 2016 in the United States, total payments for healthcare, long-term care, and hospice services for people 65 years or older with dementia were estimated to be $230.1 billion, and caregivers provided 18.2 billion hours of unpaid assistance (Alzheimer's Association 2016). The lack of effective interventions for neuropsychiatric illness is partially due to limited understanding of underlying mechanisms, but also due to under-distribution of medications and human resources in low- and middle-income countries, in which disease burden measured in DALYs is disproportionately high (Collins et al. 2011).

Autonomic nervous system (ANS) dysfunction occurs in neuropsychiatric illness, resulting in dysregulated heart rate (HR), heart rate variability (HRV), galvanic skin response, skin conductance and temperature, and respiratory rate (Draghici et al. 2016; Karemaker

2017). Due to the prevalence of HR sensors in wearable devices, and a substantial amount of literature exploring HRV measurements as markers of ANS modulation, we review studies that utilize HR and HRV measurements. Note HRV is not one metric; rather, it encompasses several types of metrics such as time domain (Stein et al. 1994; Kleiger et al. 2005; Bauer et al. 2006b), frequency domain (Akselrod et al. 1981; Montano et al. 2009), and complexity measures such as entropy (Costa et al. 2002). Changes in these metrics have been reported in patients with stress (Thayer et al. 2012), major depressive disorder (MDD; Kemp et al. 2010), bipolar disorder (BD; Henry et al. 2010), schizophrenia (Chang et al. 2009), post traumatic stress disorder (PTSD; Liddell et al. 2016), Alzheimer's disease (Femminella et al. 2014) and Parkinson's disease (PD; Maetzler et al. 2013).

Neuropsychiatric illness is also associated with alterations in behavior, especially physical movements and social routine. Patients with MDD, BD, or schizophrenia can be significantly more sedentary than age- and gender-matched healthy controls (Vancampfort et al. 2017). Diminished motor function, the presence of tremor, and coordination issues also occur in movement disorders such as Parkinson's disease. On the other hand, locomotor agitation can be a sign of mania or psychosis which may be part of the presentation of schizophrenia or BD. These abnormalities are detectable by smartphones and wearable devices with accelerometers or global positioning system (GPS) sensors. Modern smartphones and most wearables marketed to consumers for fitness purposes have accelerometers and have been explored in literature on sensing for healthcare. Behavior can also be inferred from social activity data, such as phone calls, text messages, social media use, and web browser history.

Importantly, passive monitoring via digital sensors can yield information about a patient's physiology and behavior in the 99% of the time they are not seeing a clinician, during which their symptoms fluctuate, they take actions, and are influenced by their environment in ways that profoundly impact their health (see Table A2 for example aberrations in physiology and behavior associated with illnesses; Asch et al. 2012). By shifting data from once-per-several-month visits to near real-time high frequency and high definition, improved monitoring tools

could provide a richer understanding of the day-to-day variability of neuropsychiatric illness, enable assessment of patient status before (rather than after) symptoms reach a level warranting intervention, and reduce biases and inaccuracy intrinsic in subjective questionnaires (Karow et al. 2008; Copeland et al. 2017).

Monitoring is just one step in the virtuous cycle by which data is used to improve patient management by informing interventions and resource allocation, such as home nursing visits, adjustments of a mood stabilizing medication, or telepsychiatric counseling sessions. In turn, the effect of these interventions can be measured at a hitherto unprecedented frequency and fidelity. Advances in sensor technology and informatics approaches that maximize how data is collected, analyzed, managed, and utilized by clinicians are shaping the future of care delivery in many fields of medicine. Furthermore, technology may improve the distribution of limited provider resources, especially in rural and developing parts of the world, and broader cost-effectiveness of care. Realizing this vision requires addressing many challenges, spanning technical, cultural, and economic domains. This thesis contributes technical advances to the efforts of gathering and analyzing HR and activity data via signal processing, machine learning, and information theoretical techniques.

Digital sensors in smartphones and wearables generate a vast amount of objectively measured, high-frequency, high-dimensional time series data. These data contain information about dysregulation of the ANS, social routine, and other biological rhythms (Johnson et al. 2016). In contrast, the data used in current clinical practice and biomedical research – which include self-reported symptoms, lab tests, and vital signs – are subjective, infrequently sampled, and unidimensional. Traditional methods of analysis used for these data include univariate significance testing, regression models, and simple summary statistics. However, means or medians of HR or activity may not differ in a patient with neuropsychiatric illness compared to a healthy control. Collapsing a time series into one statistic results in the loss of information related to pathophysiology.

To better capture this information, approaches from signal processing, information the-

ory, and complexity science are needed. More nuanced "features" of HR and activity time series include power spectral density attributes, wavelet coefficients, entropy (a measure of regularity or surprise), autocorrelation, or nonstationarity. These features (also called predictors or covariates) can be used to train machine learning algorithms that perform regression, continuous parameter prediction, and classification of outputs such as disease phenotype or questionnaire score (Obermeyer et al. 2016).

Machine learning algorithms estimate rules governing associations between input features and output labels without being explicitly programmed. Compared to traditional statistical methods, machine learning can utilize a large number of features even relative to the number of subjects and combine them in a nonlinear, interactive, and hypothesis-free ways. Such approaches have been applied to several problems with biomedical relevance such as detecting atrial fibrillation via a smartwatch (Tison et al. 2018), identifying diabetic retinopathy from retinal fundus images (Gulshan et al. 2016), and classifying skin cancer on par with dermatologists from dermoscopic images (Esteva et al. 2017). Overfitting is a serious limitation that occurs when an algorithm fails to generalize, i.e. can only accurately classify inputs from data used to train the algorithm, but fails to achieve high performance when presented with novel input from an external set of data not used for training. Generalizability is an obvious and crucial consideration addressed in most notable studies of machine learning applied to biomedicine via the use of cross-validation and testing on a held out set of independent or prospectively collected data.

Of note, univariate statistical significance does not guarantee predictivity or clinical utility of a biomarker (Lo et al. 2015). Methods focusing on P-values can miss useful "weak features" – those that do not significantly differ by output class when assessed via univariate statistical tests, yet can be used as input to train a multivariate machine learning algorithm that achieves high accuracy.

1.2 Opportunities to improve data collection

Digital sensors measure observations about physiology and behavior comprised of at least three components: *true signal* that reflects a clinically relevant aberration due to illness, *contextual signal* that is attributable to factors unrelated to neuropsychiatric illness, e.g. heart rate and locomotor activity patterns appear abnormal due to a temporary change in a patient's work schedule rather than a change in depressive symptoms, and *noise* or other technical challenges including insufficient sampling frequency, a lack of standardization and calibration of sensors, and noise. Maximizing true signal over contextual signal or noise, whether in the raw data itself or in the metric derived from the data, is important for improving classification of illness.

Metadata may enable the estimation of confidence in a signal. For example, GPS readings and integration with social media could capture context that discriminates if a high locomotor activity entropy is due to a subject's participation in a social gathering, cultural event, or travel for leisure as opposed to an abnormal deviation from a normal social and commuting routine. Information is not evenly distributed amongst collected data, and using complementary data sources can reduce contextual noise. The concept of data fusion and robust estimation from noisy data sources has been explored in the setting of electrocardiography (Li et al. 2008; Clifford et al. 2012a). Evaluating ambulatory HR and locomotor activity time series at specific times selected in a principled fashion could improve the richness of clinically relevant information encoded in extracted features, and improve the performance of classification tasks.

Digital sensors allow for measuring physiology at or above the Nyquist rate, which is twice the maximum component frequency of the function being sampled (Wescott 2010). This avoids the common issue of aliasing, whereby distortion or artifact occurs when the signal reconstructed from samples is different from the original continuous signal (Clifford et al. 2012b). A sampling rate of 3-6 Hz for heart rate and 10 Hz for movement is usually

sufficient for satisfying the Nyquist criterion (Winter et al. 1972; Clifford 2002). However, manufacturers of consumer devices may prioritize battery life over sampling frequency, and the latter attribute may not be reported in product documentation. Whenever possible, accounting for sampling frequency and other parameters of signal processing can prevent errors such as aliasing and improve the rigor and correctness of subsequent analysis.

Lack of standardization and calibration hinders comparisons across studies, and more importantly may limit generalizability of approaches to populations that use different technologies. Hundreds of different smartphones and wearable devices house different combinations of sensors, CPU, GPU, and operating systems.

Noise in ECG data is due to poor contact between the electrode and the skin, patient movement, muscle activity, or power line interference (Clifford 2002). Accelerometer recordings can be noisy due to thermal energy, mechanical vibrations, and the location and manner in which the device is worn (Cemer 2011). Estimation of signal quality indices and data fusion approaches can be used to detect poor quality ECG data, and these methods may also be applicable to other types of digital sensor data (Clifford et al. 2011; Clifford et al. 2012a).

1.3 Capturing information over different time scales

Interactions in a biological systems manifest in different ways over different spatial and time scales (Ivanov et al. 1999). For example, time series of BP can exhibit oscillations on the order of seconds (due to the variations in sympathovagal balance), to minutes (as a consequence of infection, blood loss, or behavioral factors), to hours (circadian rhythms and social routines) (Mancia 2012; Parati et al. 2015).

Capturing the multiscale nature of these interactions can add additional predictive information when attempting to classify illness and/or discriminate between healthy and unhealthy individuals. Costa et al. 2002 analyzed sample entropy (a metric of information complexity) of HR time series. The entropy curve, or trajectory of entropies plotted against time scale or number of coarse-grainings performed on the original time series, was found

to distinguish atrial fibrillation (AFib) from congestive heart failure (CHF) from healthy controls. Importantly, at a particular time scale the entropy values were similar for patients with AFib and for healthy controls. Likewise, there existed one time scale at which entropy was the same for AFib and CHF, and again for CHF and healthy controls.

Prior work by our group has demonstrated the additional predictive utility of analyzing HR and locomotor activity data over multiple time scales in patients with schizophrenia (Osipov et al. 2015). It is reasonable to hypothesize features of HR and locomotor activity over multiple time scales has predictive utility in other mental illnesses involving dysautonomnia. However, no work has explored the concept of multiscale dynamics in a mental health population outside of schizophrenia. Furthermore, previous explorations of multiscale dynamics in cardiovascular illness have only compared univariate measures using simple tests of statistical significance, rather than utilizing more sophisticated machine learning to perform prediction or classification that may be more clinically useful.

1.4 Interaction between time series to assess physiological or behavioral coupling

The use of multiple data streams – also known as "data fusion" – can improve classification, estimation, or prediction performance by providing a learning algorithm with non-collinear data that may contain complementary information about the underlying dynamical system, e.g. deriving many features from solely heart rate data could result in colinear features. One data stream can also be used to calculate signal quality indices for the other data (Clifford et al. 2011), and improve the accuracy of estimations of a signal in the presence of noise and other artifact (Li et al. 2008). Since smartphones and wearables contain multiple types of sensors, many monitoring approaches for neuropsychiatric illness have utilized feature extraction from multiple data streams, and data fusion at the level of the classifier. However, most studies merely present simple features from different signals to a classifier, rather than use more robust approaches that can improve the utility of information within the data, such

as signal quality indices or Kalman filters.

Transfer entropy, mutual information, and other measures of coupling assess information transfer between variables. These measures from the field of information theory have been derived from noninvasively monitored physiological data streams in efforts to characterize the dynamics of complex systems such as aging-related changes in contribution of respiration and blood pressure to entropy of heart rate (Nemati et al. 2013), and the coupling of HR and respiration in the study of respiratory-related chemosensitivity (Lee et al. 2012). However, no study to date involving multiple data streams has utilized the transfer of information between data streams collected by smartphones or wearable devices – such as HR and locomotor activity – in the context of monitoring neuropsychiatric illness. Such features may provide further predictive information about the outcome of interest compared to using features about each variable independently. A principled comparison of predictive performance when using features from just one signal, versus multiple signals, versus interactions between various data streams could improve understanding of how neuropsychiatric illness manifests in physiology and behavior, as well as further efforts to translate research studies of new technological approaches into clinically relevant advances in patient monitoring. Finally, as alluded to in the the previous section, measures of interactions between time series over multiple time scales may reflect different aspects of the perturbed physiological system than interactions assessed over solely the original time scale, and thus contain additional predictive information that enables patient classification.

1.5 Approach of thesis

This work addresses some of the problems of data collection and interaction between multiple data streams. Specifically, this thesis defends the claim that discrimination of mental illness using supervised learning algorithms is improved by considering information in HR and/or locomotor activity data during specific times (as a proxy for context), over several time scales, and between signals.

This thesis is comprised of four parts: 1) a review of past work, 2) classification of mental illness using features from quiescent segments of HR, 3) classification of mental illness using features from both HR and locomotor activity time series over varying time scales, and 4) evaluation of coupling and interactions between HR and activity as features for classifying illness.

In Part 1), the body of work upon which this thesis builds is summarized in a review of digital sensors for neuropsychiatric illness. First, the two specific mental illnesses of focus are discussed: schizophrenia and PTSD. Heart rate variability (HRV) and locomotor activity, as well as relevant metrics and features therein are reviewed. The growing literature on digital sensors for monitoring neuropsychiatric illnesses is surveyed, with a focus on passive monitoring and analyses of HR and locomotor activity, feature extraction, and classification or regression of clinically relevant outcomes (Table A4).

In Part 2), features from HR data are used to train a classifier to distinguish PTSD from control subjects. Classification is improved by isolating data from quiescent segments of HR. This work explores the hypothesis that data from quiescent segments will be more useful than data from other segments during the 24-hour recording. Sleep minimizes exogenous sources of HRV such as social routine and physical activity (Clifford et al. 2004). Dysautonomia detectable via alterations in HRV measures such as LF and HF power may thus be amplified during these quiescent segments.

In Part 3), features from both HR and locomotor activity data are used to train a classifier to distinguish subjects with schizophrenia from healthy controls. The time scale (e.g. window length) of data is varied and found to affect classifier performance. This work explores the hypothesis that information relevant to pathologically altered physiology and behavior varies with time scale.

In Part 4), interactions between HR and locomotor activity are evaluated using information theoretical approaches, and found to contribute significantly to the classification of schizophrenia compared to not using interaction features, and differently to the classification

of mental versus cardiovascular illness. This work explores the hypothesis that information between signals is altered in mental illness and relatively less altered in cardiovascular illness, and that this information is useful in a machine learning approach to discriminate patients from controls.

1.6 List of publications

The following works have been published in or accepted to a peer-reviewed journal:

- Reinertsen E, Clifford GD. A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses. Physiological Measurement. 2018;39(5):1-38. *Entirety of Chapter 2 is from this manuscript.*
- Reinertsen E, Nemati S, Vest AN, et al. Heart rate-based window segmentation improves accuracy of classifying posttraumatic stress disorder using heart rate variability measures. Physiological Measurement. 2017;38(6):1061-1076. *Entirety of Chapter 3 is from this manuscript.*
- Reinertsen E, Osipov M, Liu C, Kane JM, Petrides G, Clifford GD. Continuous assessment of schizophrenia using heart rate and accelerometer data. Physiological Measurement. 2017;38(7):1456-1471. *Entirety of Chapter 4 is from this manuscript.*
- Liu C, Oster J, Reinertsen E, Li Q, Zhao L, Nemati S, Clifford GD. A comparison of entropy approaches for atrial fibrillation discrimination. Physiological Measurement. 2018;39(7):074002. *Manuscript content referenced in Chapter 2, Section 3.3.*

The following works are submitted to a peer-reviewed journal or in preparation:

- Reinertsen E, Shashikumar SP, Nemati S, Clifford GD. Multiscale network dynamics between heart rate and locomotor activity are altered in schizophrenia. Submitted to Physiological Measurement. *Entirety of Chapter 5 is from this manuscript.*
- Cakmak A, Reinertsen E, Nemati S, Clifford GD. Benchmarking changepoint detection algorithms on cardiac time series. In preparation. *Manuscript content referenced in Chapter 6, Section 3.1.*

1.7 Contributions from others

Gari D. Clifford provided intellectual guidance and contributed substantially to algorithmic and experimental design for all work in this thesis.

Qiao Li wrote code for signal quality indexing and pulse detection from PPG waveforms.

Shamim Nemati and Joon Lee wrote code for performing Darbellay-Vajda partitioning and calculating transfer entropy. Supreeth Shashikumar and Shamim Nemati wrote code for transforming time series into networks and for analyzing multiscale network representation features.

Amit J. Shah collected and provided data from the PTSD patient cohort. Rachel Lampert analyzed some of these data and provided HRV measures.

John M. Kane, Georgios Petrides, and Yashar Behzadi collected and provided data from the schizophrenia patient cohort. Maxim Osipov provided code for loading these data and performing initial data quality checks.

Ayse Cakmak modified the Bayesian change point detection algorithm and wrote code to compare various change point detection methods.

CHAPTER 2

DIGITAL SENSORS FOR NEUROSPSYCHIATRIC ILLNESS

2.1 Overview

This chapter provides an overview of two mental illnesses the work focuses on – schizophrenia and PTSD. HRV and locomotor activity are discussed. Finally, relevant studies of the aformentioned illnesses that use digital sensors are reviewed to provide a foundation of previous work in the field, organized by major devices: smartphones, wearable devices, and Holter monitors.

2.2 Mental illness

2.2.1 Schizophrenia

Schizophrenia is a complex and heterogeneous psychiatric disorder characterized by several criteria, including at least two of the following symptoms for one month or longer: delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behavior, or negative symptoms such as diminished emotional expression; furthermore, there must be impairment in work, interpersonal relations, or self-care, as well as continuous signs of the disorder for at least 6 months (American Psychiatric Association 2013). The DSM-V criteria are provided in Table 2.1. The lifetime global prevalence of schizophrenia is about 1%, and numerous risk factors include complications during fetal life, older paternal age, male gender, being in a disadvantaged inner city environment, migrant status, cannabis use, and childhood adversity (Kahn et al. 2015). The mechanisms of pathophysiology are far from fully understood, with evidence of abnormal neurotransmitter signaling and receptor function, reductions in grey and white brain matter, and complement-mediated synapse elimination during postnatal development. Outcomes vary widely, ranging from total recovery to totally debilitating

illness requiring chronic care. Life expectancy for people with schizophrenia is reduced by 20 years compared to people without the illness. Pharmacological treatments for schizophrenia can relieve psychotic symptoms but usually fail to meaningfully improve social, cognitive and professional functioning. Psychosocial interventions can be useful but are resource-intensive and inconsistently delivered. Finally, schizophrenia tends to be diagnosed years after symptoms begin. Relatively little work at the intersection of mental health and digital sensing technology has focused on schizophrenia compared to depression, BD, or dementia.

Table 2.1: DSM-V criteria for schizophrenia[†]

Criterion	Description
Criterion A	Two or more of the following symptoms for > 1 month unless treated successfully include delusions; hallucinations; disorganized speech; disorganized or catatonic behaviour; and negative symptoms, such as affective flattening or loss of initiative
Criterion B	Level of functioning is significantly decreased in work, personal relationships and/or personal care
Criterion C	Symptoms of the disorder last $\geq$6 months
Criterion D	Exclusion of schizo-affective disorder, unipolar and bipolar affective disorder
Criterion E	Symptoms cannot be attributed to the use of drugs or medication, or to a somatic disorder
Criterion F	In the case of a pre-existing autism spectrum disorder, at least 1 month with prominent hallucinations or delusions

[†]American Psychiatric Association 2013

2.2.2 Post traumatic stress disorder

PTSD is a psychopathological response that can develop after exposure to traumatic events such as violence, natural disasters, or combat. It affects multiple systems ranging from brain chemistry and circuitry to cellular, immune, metabolic, and endocrine function. The DSM-V diagnostic criteria for PTSD are shown in Table 2.2.

Symptoms can include nightmares of the trauma, hypervigilance, difficulty sleeping, poor concentration, and avoidance of places, activities, or persons that remind the affected individual of the causal incident (Yehuda et al. 2015). PTSD has a lifetime prevalence of about 8% in the US general population (Resnick et al. 1993). The prevalence of PTSD is higher

Table 2.2: DSM-V criteria for PTSD[‡]

Criterion	Description	Specific examples	Requirements
Criterion A	Exposure to stressor	• Direct exposure • Witnessing trauma • Learning of a trauma • Repeat or extreme indirect exposure to aversive details	Exposure to trauma can occur either by direct or indirect confrontation with extreme trauma
Criterion B	Intrusion symptoms	• Recurrent memories • Traumatic nightmares • Dissociative reactions (flashbacks) • Psychological distress at traumatic reminders • Marked physiological reactivity to reminders	At least one of these five examples is required
Criterion C	Persistent avoidance	• Trauma-related thoughts or feelings • Trauma-related external reminders such as people, places or activities	At least one of these two examples is required
Criterion D	Negative alterations in cognitions and mood	• Dissociative amnesia • Persistent negative beliefs and expectations • Persistent distorted blame of self or others for causing trauma • Negative trauma-related emotions: fear, horror, guilt, shame and anger • Diminished interest in activities • Detachment or estrangement from others • Inability to experience positive emotions	At least two of these seven examples are required
Criterion E	Alterations in arousal and reactivity	• Irritable and aggressive behavior • Self-destructive and reckless behavior • Hypervigilance • Exaggerated startle • Problems concentrating • Sleep disturbance	At least two of these six examples are required
Criterion F	Duration	Must experience criteria B, C, D and F for > 1 month	Acute stress disorder is diagnosed for symptoms occurring for < 1 month post trauma
Criterion G	Functional significance	Impairment in social, occupational or other domains	Disability in at least one of these domains is required
Criterion H	Exclusion	Not attributable to medication, substance use or other illness	Symptoms must not be secondary to other causes

[‡]American Psychiatric Association 2013

in developing or war-afflicted countries, in which people are exposed to more severe and/or more numerous traumas (Karam et al. 2014). Lifetime prevalence is thus especially high in veterans, ranging from 6-30% (Dohrenwend et al. 2006; Kok et al. 2012; Sundin et al. 2014; Marmar et al. 2015).

2.3 Heart rate variability

Dysfunction of the balance between sympathetic and parasympathetic branches of the ANS can manifest as differences in HRV – beat-to-beat variation in heart rate. Although not routinely used for diagnosis or monitoring, HRV is altered in neuropsychiatric illnesses including depression, anxiety, schizophrenia, and PTSD (McCraty et al. 2001; Agelink et al. 2002; Cohen et al. 2000; Beauchaine 2001; Alvares et al. 2016; Draghici et al. 2016). The relationship between vagal tone and behavioral reactivity, psychological status, or psychiatric disease is incompletely understood as it changes with age, is confounded by illness such as cardiovascular disease, and lacks mechanistic explanations of underlying pathophysiology. Complicating matters, commonly used anti-psychotic medications affect the ANS; subsequent alterations in HRV may provide a way of measuring adherence to pharmacological therapy (O'Regan et al. 2015). HRV can be monitored non-invasively using wearable technology such as photoplethysmography on a smart watch, or an electrocardiogram (ECG) on an adhesive patch. Several categories or types of HRV measures are summarized here and include time domain, frequency domain, and entropy measures. For a more comprehensive description of HRV measurements and the literature exploring clinical utility beyond neuropsychiatric illness, we refer the reader to the excellent review by Kleiger et al. 2005.

2.3.1 Time-domain HRV metrics

In time domain analysis, the intervals between adjacent normal R waves are characterized, and a variety of descriptive statistics can be computed from the intervals directly, or the differences between the intervals.

SDNN is the standard deviation of all normal RR intervals, and the most commonly reported time domain HRV measure. 30–40% of SDNN magnitude is attributable to day:night difference in NN intervals. The SDNN is sensitive to ectopic beats, artifacts, and missed beats (Clifford 2006). These events can artificially shorten or elongate RR intervals and thus affect SDNN. At least 20 hours of ambulatory monitoring recordings may be required to calculate SDNN (or other time- and frequency domain measures) (Haaksma et al. 1998).

SDANN is the standard deviation of 5-minute average NN intervals, which provides a smoothed version of SDNN reflecting long-term fluctuations (Bigger et al. 1989). SDANN is less sensitive to errors caused by individual artifacts than SDNN because "averaging many NN intervals minimizes the effects of unedited artifacts, missed beats, and ectopic complexity" (Kleiger et al. 2005).

ASDNN, or SDNN index, is the mean of the standard deviations of all NN intervals for all 5-minute segments in 24 hours. This metric significantly correlates with SDNN and SDANN because low and high HRV tend to be global phenomena (Kleiger et al. 1992).

rMSSD is the square root of the mean of the squares of successive NN interval differences. This metric reflects the average change in interval between beats (Pagani et al. 1985).

NN50 is the number of NN intervals differing by > 50 ms from the preceding interval (Mietus et al. 2002). pNN50 is the percentage of intervals > 50 ms different from the preceding interval, or NN50 normalized by the total number of RR intervals. In the presence of normal sinus rhythm and normal AV-nodal function, each of these measures quantifies parasympathetic modulation of normal RR intervals driven by ventilation.

Phase-rectified signal averaging

The fast Fourier transform assumes that the signal is stationary and continues for infinite time. However, RR interval time series exhibit non-stationarity, and are finite in length. If the signal consists of many short patches of periodicities with a particular frequency, some of the patches will be in phase with the analyzing sinusoid, while most will be out of

phase. Patches with phase shifts differing by π will cancel; only the few patches with no corresponding patch with a phase shift of π will contribute to the Fourier coefficient and subsequent power spectra.

Phase-rectified signal averaging (PRSA) is a time-domain method developed to compress a time series signal into a shorter sequence, keeping relevant quasi-periodicities but eliminating non-stationarities, artifacts, and noise (Kantelhardt et al. 2007). First, the signal is re-sampled evenly in time. Next, anchor points, or some of the indices i, are selected according to an increase, a decrease, or a change in mean in the signal. For each anchor point, a window ranging from $i-L$ to $i+L$ is defined where L is the distance from the index to the boundary of the window. Every window in the time series is aligned and averaged to generate a PRSA transform, or "curve" $\bar{x}(k)$ where k is the offset from the anchor point. Because anchor points are phase synchronized with the signal, all patches contribute to the PRSA signal and its power spectra regardless of non-stationarities.

One useful property of the PRSA is that the amplitude of a spectral component at frequency f is determined by $A_f^2 f$, whereas the amplitude of the original signal from which the PRSA is calculated is given by A_f (Bauer et al. 2006b). The power spectrum of a signal, $P(f)$, is proportional to the square of the amplitude. For $\frac{1}{f}^{\beta}$ noise with a scale coefficient β, the power spectrum is given by $P(f) \sim A_f^2 \sim f^{-\beta}$. We derive the expression for the power spectrum of the PRSA, given by $P_{\mathrm{PRSA}}(f) \sim p_f^2 \sim A_f^4 f^2 \sim f^{-2\beta+2}$. Thus, transforming a signal via PRSA and assessing its power spectrum results in a larger scale coefficient that is easier to detect when searching for deviations from standard scaling behaviour that are caused by quasi-periodicities.

The PRSA curve is quantified by estimation of central wavelet coefficients $\tilde{x}_w(s,0)$ at various scales s. Examples of continuous common wavelets include derivatives of Gaussians such as $g_1(t) = t\exp(-t^2/2)$ (first derivative), 'Mexican hat' wavelet $g_2(t) = (t^2-1)\exp(-t^2/2)$ (second derivative), and the Haar wavelet $h(t) = 1$ for $-1 \leq t < 0)$, $+1$ for $0 \leq t < 1)$, 0 (else).

Deceleration capacity (DC) determines the capacity of the central nervous system to quickly decelerate the rate. This HRV measure uses the Haar wavelet, $s = 2$, and the anchor point definition of increasing RR intervals (decreasing heart rate), and is given by $\tilde{x}_h(2, 0)$. DC has been shown to be a better predictor of mortality in survivors of myocardial infarction than left ventricular ejection fraction (LVEF), the current gold standard risk predictor (Bauer et al. 2006a; Kantelhardt et al. 2007). Additionally, in a study of PRSA metrics evaluated in subjects with schiozphrenia, DC has been shown to be reduced in subjects taking antipsychotic medication (Birkhofer et al. 2013). Acceleration capacity (AC) - similar to DC except with an anchor point definition of decreasing rather than increasing RR intervals - and DC may be sensitive markers for autonomic changes associated with aging (Campana et al. 2010).

2.3.2 Frequency-domain HRV metrics

Power spectral density (PSD) analysis via fast Fourier transformation or autoregression techniques quantifies the frequency components of RR intervals (Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology 1996; Clifford 2002). The PSD is partitioned into very low frequency (VLF) <0.04 Hz, low frequency (LF) 0.04 - 0.15 Hz, and high frequency (HF) >0.15. LF power is modulated by baroreflexes, sympathetic, and parasympathetic tone (Billman et al. 1990; Furlan et al. 1990; Bloomfield et al. 1997), whereas HF power is modulated by parasympathetic activity (Katona et al. 1970; Appel et al. 1989; Billman et al. 1990; Thayer et al. 2010). Pagani et al. 1985 proposed that the LF/HF ratio quantifies the relationship between sympathetic and parasympathetic tone, which assumes a linear interaction between the two branches of the ANS on HRV. However, the assumptions underlying this simple model coined "sympathovagal balance" have been challenged by evidence showing complex nonlinear effects of varying cardiac sympathetic and parasympathetic nerve activity on LF/HF; some reports suggest sympathetic activity may modulate HF power, and conversely parasympathetic activity may

modulate LF power (Billman 2013).

Removing noise and ectopic beats from RR interval data is necessary for the calculation of HRV measures. The frequency of ectopic beats and artifacts are linearly related to the error of the PSD estimate (Clifford et al. 2005). Furthermore, because RR intervals are irregularly spaced in time, resampling and interpolation of data into uniformly spaced intervals is required to subsequently perform a Fourier transformation. However, resampling can result in over-estimation of the PSD. The resampling period affects the computational speed and error of PSD estimation; Singh et al. 2004 proposed a now widely used sampling frequency of 4 Hz for the study of autonomic regulation, since it enables the computation of spectral estimates between DC and 1 Hz which represents an adequate range of autonomic nervous system response. The Lomb-Scargle periodogram is a more appropriate spectral estimation technique for unevenly sampled time series, as it provides a less noisy estimate of the PSD compared to other methods such as the Welch periodogram. Clifford *et al.* has shown it is superior to other spectral estimation methods for HRV interval data (Clifford et al. 2005).

The segment length of RR recordings is also an important consideration for discrete Fourier transform analysis and is governed by the compromise between the need for short data segments to give acceptable variance and stationarity considerations, and the need for long data length to give acceptable frequency domain resolution and reduced spectral leakage (Singh et al. 2004). The recording should be at least 10 times the lower frequency bound of the investigated component, but no longer to reduce the probability of nonstationarity (Clifford 2002). Commonly reported segment lengths are 5 minutes and 24 hours. A shorter recording is more feasible from a research logistics standpoint and will capture high frequency components. However, short recordings are likely to undersample rare events such as atrial fibrillation, or phenomena that occur over longer time scales such as changes in HRV due to the circadian rhythm.

In practice, studies use a wide range of resampling rates, window types, overlap between segments, and RR interval segment lengths (Singh et al. 2004). In an effort to improve

the standardization, transparency, and reproducibility of HRV analyses, Vest et al. 2017 developed an open-source HRV analysis toolbox in the Matlab language. The toolbox uses the minimal number of dependencies and the most basic operators to future-proof the code.

2.3.3 Entropy

In information theory, entropy is defined as the expected value of information in a message. Signals such as a sine wave or white noise contain little information, and accordingly have low entropy. On the other hand, the human heart beat is richly complex, with dynamics influenced by multiple physiological control systems acting over multiple time scales, and has higher entropy. Entropy and other conceptually related measures calculated from RR interval time series can quantify the regularity or complexity of underlying physiology and can reflect alterations in ANS function due to neuropsychiatric illness (Osipov et al. 2015). This and other measures can be useful in distinguishing time series with similar mean μ and variance σ values as well as similar power spectral densities, albeit differing levels of complexity (Figure 2.1).

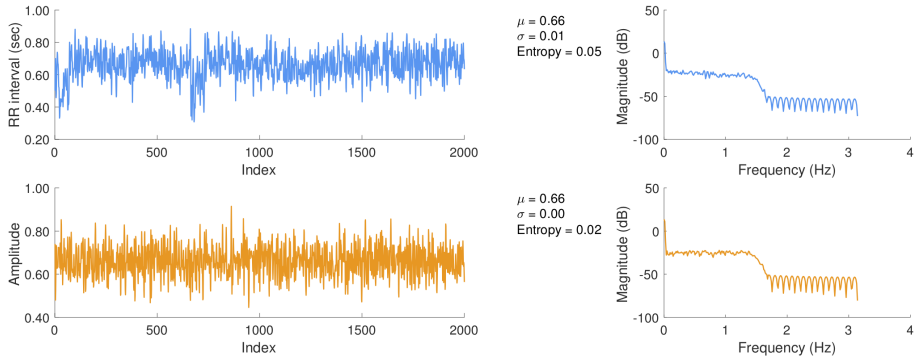


Figure 2.1: Two toy signals (blue in upper left and orange in lower left) with similar means μ, variances σ, and power spectra (right upper and right lower subplots). Note the blue signal has an entropy of 0.05, whereas the orange signal has an entropy of 0.02, indicating different complexities.

Sample entropy

Sample entropy (SampEn) H, first defined by Richman et al. 2000, is a metric of signal complexity derived from the negative logarithm of the conditional probability of the appearance of longer patterns in a signal, considering the presence of a shorter pattern:

$$H(m, r, N) = -\ln \frac{A^m(r)}{B^m(r)} \tag{2.1}$$

where m is the template length, r is the radius of similarity or distance threshold between patterns (normalized to be unitless), $A^m(r)$ is a probability of matching a template of length $m+1$, $B^m(r)$ is the probability of matching a template of length m, and N is the number of elements in the time series (not explicit in the expression, but affects the final value of H). Two patterns of length m are considered similar if each point of a pattern in one part of the signal is within distance r from the respective point in the other part of the signal. Because $A^m(r) \leq B^m(r)$, SampEn is ≥ 0. For a finite N, the theoretical upper bound of SampEn is $\leq \ln(N - m)$. (Richman et al. 2000).

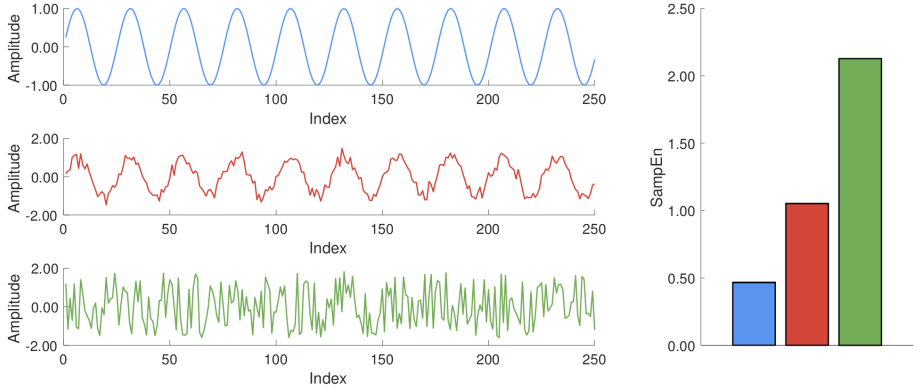


Figure 2.2: Three signals with progressively increasing complexity quantified by SampEn. The blue signal in the upper left subplot is a sine wave with minimal complexity, indicated by a SampEn of 0.50. The red signal in the middle left subplot is a sine wave with added noise and thus additional complexity, indicated by a higher SampEn value around 1.00. The green signal in the lower left subplot is generated by superimposing Gaussian noise on a sine wave with the same phase and amplitude as in the previous two plots, and thus has the highest complexity, indicated by the SampEn value above 2.00.

Lanata et al. 2015 evaluated a small cohort of ten bipolar patients using shirts with em-

bedded ECG sensors. Subjects were administered questionnaires via a smartphone application. Increased SampEn of heart rate time series was associated with clinical improvements defined by mood transitions from a pathological mood state (featuring depressive and/or hypomanic symptoms).

Multiscale entropy

To assess complexity at different time scales, the original signal can be coarse-grained to a lower sampling frequency. Specifically, data points within non-overlapping windows of increasing length τ are averaged. For the τ_{th} time scale, each element of the coarse-grained time series, $y_j^{(\tau)}$, is given by:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i \tag{2.2}$$

where τ represents the scale factor and $1 \leq j \leq N/\tau$. The length of each coarse-grained time series is N/τ. The first time scale corresponds to the original time series, the second time scale corresponds to one coarse-graining, etc. Another method of coarse-graining is to sample every i^{th} data point, but by ignoring all other data points results in more loss of information compared to the first method. Entropy or other complexity metrics can then be calculated using $y_j^{(\tau)}$. MSE of actigraphy time series data was a predictive feature that discriminated patients with schizophrenia from controls (Osipov et al. 2015).

Fuzzy entropy

Although SampEn differs by illness status, it is sensitive to small changes in parameter values and thus exhibits poor statistical stability. Additionally, SampEn only accounts for similar patterns with similar amplitudes, not similar patterns with different amplitudes. These shortcomings are addressed by the recently proposed fuzzy entropy $\mathcal{H}_{\text{fuzzy}}$ (Liu et al. 2013).

To calculate $\mathcal{H}_{\text{fuzzy}}$, the binary Heaviside classifier in SampEn is replaced with a continuous membership degree between 0 and 1, based on Zadeh's concepts of fuzzy set theory.

$\mathcal{H}_{\text{fuzzy}}$ is more robust to noise than SampEn and is calculated in a similar manner as MSE. The k'th MFE value is given by $\mathcal{H}_{\text{fuzzy}}(y^{(k)})$ where $y^{(k)}$ is the original time series x after k coarse-graining steps.

Fuzzy entropy also normalizes the amplitude of signals by subtracting the local or global mean or maxima, which enables comparison of signals by trend rather than amplitude alone. Previous entropy measures would fail to consider two sequences as a match if they differed in amplitude by more than r even if the trend was identical. For example, consider sequences $a = [7, 8, 7]$, $b = [7, 8, 8]$, and $c = [3, 4, 3]$. The Euclidean distance between a and b is given by

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} = 1 \tag{2.3}$$

whereas $d(a, c) \approx 6.93$ and $d(b, c) \approx 8.25$. Although vectors a and b are closer together in Euclidean space compared to vectors a and c, vectors a and c have a similar trend of a one-unit increase followed by a one-unit decrease. Depending on the selected radius of similarity r, previous entropy measures would consider only vectors a and b as matches, whereas $\mathcal{H}_{\text{fuzzy}}$ might also consider vectors a and c as matches. Thus, $\mathcal{H}_{\text{fuzzy}}$ could better detect temporal trends in a signal compared to the amplitude-dependent SampEn. Liu et al. 2013 demonstrated fuzzy entropy of RR interval time series better separated heart failure patients from healthy controls via a univariate significance test compared to other entropy measures.

2.3.4 Effects of medications on HRV

Antipsychotic medications exert antagonistic and agonistic activity on muscarinic and α-adrenergic receptors, and also alter ANS function. This complicates assessment of ANS dysfunction in patients with mental illness who are taking these medications. Hattori et al. 2017 investigated how four atypical antipsychotic drugs – risperidone, olanzapine, aripipra-

zole, and quetiapine – affect ANS activity as assessed via spectral analysis of HRV. 241 patients with schizophrenia received an atypical antipsychotic as monotherapy; 90 subjects received risperidone, 68 olanzapine, 52 aripiprazole, and 31 quetiapine. The quetiapine group showed significantly diminished sympathetic and parasympathetic activity compared with the risperidone and aripiprazole groups, with lower LF and HF power. Furthermore, multiple regression analysis showed that the type of antipsychotic drug significantly influenced ANS activity.

Van Zyl *et al.* reviewed the literature on how antidepressant medications affect HRV (Zyl et al. 2008). Tricyclic antidepressants (TCAs) were associated with declines in most measures of HRV and significant increases in HR in studies with short recordings (2-10 minutes). Interestingly, no significant changes were found for longer recordings (24 hours). Treatment effects with selective serotonin reuptake inhibitors (SSRIs) were more variable. Short recording studies revealed a significant decrease in HR and an increase in one HRV measure. In two 24-hour recording studies no significant changes were observed.

To account for the affects of pharmacological agents on HRV in patients with mental illness, studies should maintain subjects on the same medication regimen, and avoid starting or discontinuing a new therapy. No study has explored if noninvasive measurements of HRV, locomotor activity, or other continuously monitored data are useful in the evaluation of medication adherence status of a patient. This application is challenged by the variability of ANS response to different medications, the myriad of medications taken by HRV patients, co-existing cardiovascular conditions, as well as numerous confounding factors that affect sensor readout.

2.4 Locomotor activity and behavior

2.4.1 Rest-activity characteristics

Rest-activity characteristics can quantify variability and amplitude of locomotor activity, and include mean level during the least active five hours (L5), mean level during the most active

ten hours (M10), relative amplitude (RA), interday stability (IS) and intraday variability (IV) (Witting et al. 1990; Van Someren et al. 1999).

L5 and M10 are calculated by sliding a five or ten-hour window through all data, calculating the mean value of each window, and returning the lowest or highest value respectively. M10 is activity during the most active period of the day, whereas L5 represents activity during sleep plus nighttime arousals. RA is a nonparametric and dimensionless measure that relates M10 to L5:

$$\mathrm{RA} = \frac{\mathrm{M10} - \mathrm{L5}}{\mathrm{M10} + \mathrm{L5}} \tag{2.4}$$

The interdaily stability (IS) quantifies the invariability of activity between days. The IS is equivalent to the 24 hour value from the chi-square periodogram normalized for the number of data (Sokolove et al. 1978), and is calculated as the ratio between the variance of the average 24 hour pattern around the mean and the overall variance:

$$IS = \frac{n \sum_{h=1}^{p} (\bar{x_h} - \bar{x})^2}{p \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.5}$$

where n is the total number of data, p is the number of data per day, $\bar{x_h}$ are hourly means, $\bar{x}$ is the mean of all data, and x_i represents individual data points. IS varies between 0 for Gaussian noise and 1 for cyclical activity whereby the avareage 24 hour variance around the mean is equivalent to the overall variance.

The intradaily variability (IV) indicates fragmentation in the daily activity rhythm by quantifying the "frequency and extent of transitions between rest and activity" (Van Someren et al. 1999), and is calculated as the ratio of the mean squared first derivative and the population variance:

$$IV = \frac{n \sum_{i=2}^{n} (x_i - x_{i-1})^2}{(n-1) \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.6}$$

IV varies between 0 for a perfect sine wave to ≈ 2 for Gaussian noise.

Rest-activity characteristics have been found to differ in subjects with neuropsychiatric illnesses. Subjects with schizophrenia had higher IS and lower IV than the controls, reflecting a more structured behavioral pattern (Berle et al. 2010). This difference was even more pronounced in subjects with schizophrenia treated with Clozapine and was not found in depressed patients. Additionally, IS and IV differed among subjects with schizophrenia, depression, and controls via ANOVA.

2.5 Monitoring approaches

2.5.1 Smartphones

Smartphones are globally ubiquitous, owned by 72% of Americans and 3B people worldwide, and are projected to reach a global total of over 5B people by 2030 (Poushter 2016). Importantly, studies in the USA, United Kingdom, Canada, and India have found smartphone ownership to not be significantly lower among people with serious mental health conditions compared to the average owner, and ownership by these individuals is projected to increase, mirroring the trend seen in the general population (Torous et al. 2014; Firth et al. 2016). Additionally, people tend to keep their phones with them and check them between 46 to 85 times per day (Andrews et al. 2015; Eadicicco 2016). These data thus reflect social and behavioral manifestations of neuropsychiatric illnesses in the context of daily life rather than in an artificial clinical setting (Insel 2017). For example, GPS location data measured on smartphones can be used to estimate behavioral attributes such as percentage of time a subject spends in certain locations (Figure 2.3). By evaluating the time of day, day of week, and amount of time spent in each location, the purpose of each location datum can be in-

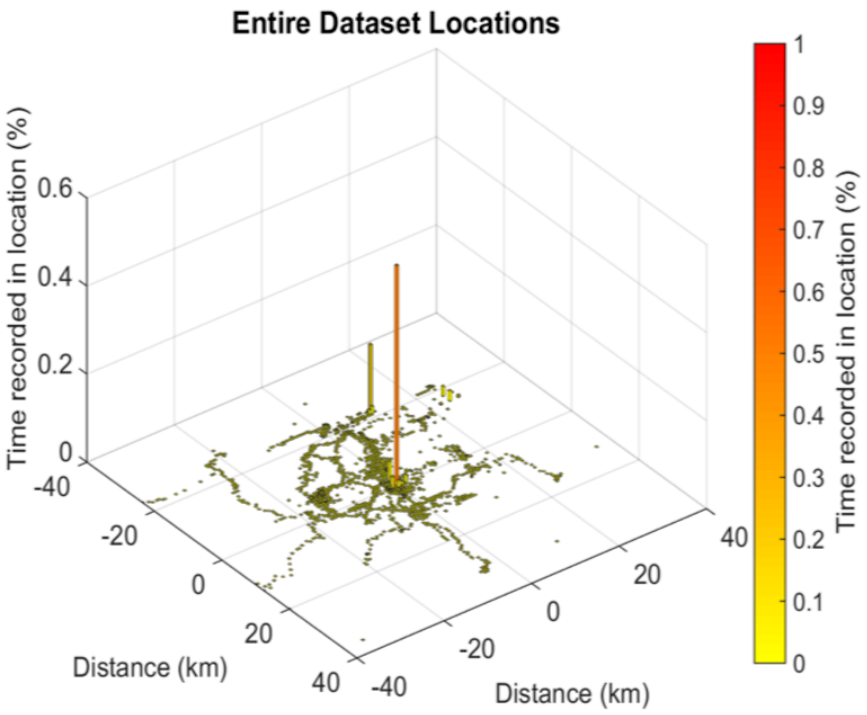
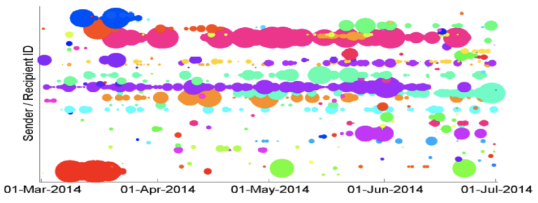


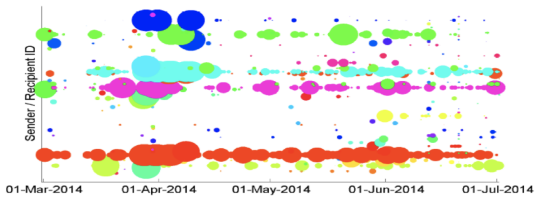
Figure 2.3: Geolocation data measured via smartphone can track time spent at modal locations. The x- and y-axes are distance from the most commonly visited location. The z-axis is the percentage of total time spent in a given location, with darker orange encoding a higher percentage and a lighter yellow encoding a lower percentage. The dark orange peak at the origin where the individual spends the most time is assumed to be home, and the second-largest peak (z-axis value) where the individual spends the next most time is assumed to be work, or vice-versa if the individual spends more time at work than home.

ferred, e.g. work versus home. Additionally, social interactions in the form of calls and text messages can be monitored and quantified (Figure 2.4). Geolocation, social network activity, and other attributes reflect behavior and may differ in subjects with neuropsychiatric illness compared to healthy controls. Several investigators have built smartphone apps for collecting sensor and usage data, including Automated Monitoring of Symptom Severity (AMoSS; Palmius et al. 2014), Purple Robot (Schueller et al. 2014), and Beiwe (Torous et al. 2016). A long-term goal for the work described in this thesis is to build a smartphone platform to enable disease monitoring and classification.

Smartphones can also be used to administer validated questionnaires for evaluating quality of life and mental well-being (Palmius et al. 20177). Although self-reported questionnaires



(a) Social network activity levels by contact (healthy control)



(a) Social network activity levels by contact (bipolar)

Figure 2.4: Social network activity measured via smartphone can identify mood and illness. The y-axis encodes unique pairings of sender and recipient IDs. The x-axis encodes time. The radius of each colored dot is proportional to the number of calls and text messages in one day. Interactions from a sender-recipient pairing have the same color over time, i.e. all red dots with the same height on the y-axis represent interactions between the same two unique individuals. Qualitatively, (a) healthy controls demonstrate more regular amounts of interaction over time with their social contacts compared to (b) subjects with bipolar disorder who alternate bouts of high and low levels of interaction.

are prone to recall, social desirability, and confirmation biases, they provide a pragmatic best estimate of an individual's mental status and can achieve results comparable to clinician-administered surveys (Spitzer et al. 1999; Ebner-Priemer et al. 2006; Martel 2008; Solhan et al. 2009). The inference of mental health questionnaire results from digital sensor data is a common approach in the literature and could be useful for monitoring the status of subjects who struggle with adherence or have impaired cognition and executive decision-making capacity (Table A3; Mohr et al. 2016; Tsanas et al. 2016; Barrett et al. 2017; Aung et al. 2017). In this section we review recent studies using smartphones to monitor neuropsychiatric illnesses; work that may be related but also involves analysis of heart rate data is reviewed in later sections.

Monitoring schizophrenia via smartphone

Wang *et al.* collected passive smartphone sensor data from 21 outpatients diagnosed with schizophrenia and recently discharged from hospital over a period ranging from 2 - 8.5 months (Wang et al. 2016). Samsung Galaxy S5 phones running the Android operating system were equipped with the "CrossCheck" app developed in-house that monitors type and duration of physical activity, sleep duration, number and durations of phone conversations,

number of SMS, geolocation, phone and app usage, and ambient light and noise. Every three days, Ecological Momentary Assessment (EMA) questions were administered and sensor data were aggregated. Generalized estimating equations were used to map associations between features and EMA responses. Higher scores in attributes related to positive perception of mental well-being – including calm, hopeful, sleeping well, social, and ability to think clearly – were associated with waking up earlier, having fewer conversations, fewer phone calls, and fewer SMS. Higher scores in questions related to negative perception were associated with staying stationary more in the morning but less in the evening, visiting fewer new places, having fewer conversations but making more phone calls and SMS, and using the phone less. Gradient boosted regression trees were used to predict EMA scores from these features. Models trained on an individual's data estimated EMA scores for that individual with a correlation between prediction and outcome of $r = 0.77$ and $p < 0.001$. However, outcomes predicted via leave-one-out cross validation did not correlate with actual outcomes. A population-wide model may be less useful for individualized predictions given the high variance of feature phenotypes, compared to models trained on each specific individual's past data.

Staples *et al.* recently reported a three-month observational study of both self-reported and objective measures of sleep in schizophrenia (Staples et al. 2017). Using the Beiwe app (available for both iPhones and Android phones), 13 subjects diagnosed with schizophrenia were given tri-weekly EMAs. Passive data were continuously collected, including accelerometry, GPS, screen use, and anonymized call and SMS activity. Sleep quality was assessed in a clinical setting using the PSQI, which was compared to both EMAs and sleep estimates based on passively collected accelerometer data. A cross-validated linear regression model with mean phone-based EMA scores as the outcome and mean paper-based PSQI scores as the predictor classified 85% (11/13) of subjects as exhibiting high or low sleep quality. Accelerometry moderately correlated with subject self-assessments of sleep duration ($r = 0.69$, 95% CI $[0.23 - 0.90]$). Active and passive phone data predicted concurrent PSQI scores with

a mean average error of 0.75, and future PSQI scores with a mean average error of 1.9, with scores ranging from 0-14.

Among individuals who are diagnosed, hospitalized, and treated for schizophrenia, up to 40% of those who are discharged will relapse within one year. Barnett *et al.* evaluated a smartphone platform for monitoring seventeen patients with schizophrenia undergoing active treatment in order to identify warning signs of relapse, defined as psychiatric hospitalization or an increase in the level of psychiatric care, such as increase in the frequency of clinic visits or referral to a partial or outpatient hospital program (Barnett et al. 2018). Patients were monitored for three months using the Beiwe app, and mobility patterns and social behavior were gathered and analyzed. Features were extracted from the data, including daily distance traveled, time spent at home, number of significant locations visited, total duration of calls, number of missed calls, and number of text messages sent. The app also administered surveys twice per week to assess anxiety, depression, sleep quality, psychosis, the warning symptoms scale, and medication adherence. The rate of behavioral anomalies detected in the 2 weeks prior to relapse was 71% higher than the rate of anomalies during other time periods. Although anomalies were calculated using each patient's own data to account for differences in baseline features, the number of anomalies greatly varied between subjects. Additionally, many subjects did not relapse, as the cohort enrolled only seventeen patients and for only three months. The features captured in patients that did relapse may not have reflected the "potential trajectories and mechanisms that can lead to relapse". The anomaly detection approach demonstrated in this paper could be useful for measuring other outcomes that were not reported but could be clinically useful, such as changes in positive or negative symptoms of schizophrenia.

Monitoring PTSD via smartphone

Smartphone apps for PTSD have focused on education about the disorder, delivery of cognitive behavioral therapy, self-assessment of symptoms via questionnaires, and access to crisis

support and other relevant resources (Kuhn et al. 2014). Few papers describe the use of digital sensors to passively monitor clinical symptoms of PTSD. However, many smartphone- and wearables-based sensing approaches have focused on anxiety and depression which are common co-morbidities.

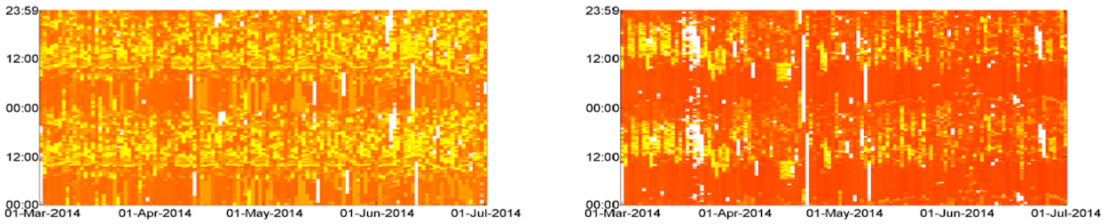
Place *et al.* conducted a 12-week trial with 73 patients who reported at least one symptom of PTSD or depression (Place et al. 2017). Clinical symptoms were assessed by licensed social workers who administered the depression and PTSD modules of the Structured Clinical Interview for Mental Disorders (SCID). An Android app was developed to gather accelerometry, SMS and call, location, device use, and audio data. Extracted features included sum of outgoing calls, count of unique numbers texted, absolute distance traveled, dynamic variation of the voice, speaking rate, and voice quality. Feature reduction was performed to reduce over-fitting and interfeature correlation, and a logistic regression was trained using 10-fold cross validation. Fatigue was not accurately predicted, with an AUC of only 0.56. Predictions of interest in activities, social connectedness, and depressed mood featured much better AUCs of 0.75, 0.83, and 0.74 respectively, which is closer to the discrimination performance of currently used psychiatric tools and other clinical assays. Finally, subjects reported comfort with sharing personal data with clinicians and medical researchers. However, it was unclear if the predictive model outperformed sample-and-hold estimations of mood from the previous week. This can be viewed using a Bayesian framework, in which the mood state from the previous week informs the prior, and data from the smartphone app is used to update the model and estimate the posterior. Evaluating subjects at several time points affords an opportunity to quantify the additional contribution of passive sensor data to predictive models that use questionnaires or surveys as ground truth.

University of North Carolina, Harvard University, and Verily Life Sciences LLC (South San Francisco, CA) are leading the AURORA study, a 19-institution five-year effort to perform the most comprehensive observational study of mental disorders that occur in the wake of trauma to date (National Institute of Mental Health 2016). Investigators will screen 5,000

people arriving in emergency rooms after trauma. After an initial evaluation and a baseline collection of biological data from blood samples, subjects will be monitored for the next several months through the use of mobile technology, such as wrist wearables and smart phones, to track factors like activity, sleep, and mood. Other assessments will include additional blood samples, functional brain imaging, and psychological tests. Participant involvement will continue over a year, generating a wide variety of detailed information on, for example, health history (including that of earlier trauma), genetics, stress responses (physical and psychological), behavior, and cognition. This collaboration presents a unique opportunity to learn more about the factors that mediate the development of mental illness after trauma, and potentially contribute to new diagnostic and therapeutic approaches. The Aurora study represents a new trend in public-private partnerships, involving multiple research institutions and and technology companies such as Verily and Mindstrong Health (Palo Alto, CA).

2.5.2 Wearable accelerometers

Locomotor activity is altered in neuropsychiatric illnesses, due to impaired motor function, weakness, volitional and behavioral changes, or abnormal sleep patterns and circadian rhythms (Teicher 1995). Non-invasive body-worn accelerometers can measure these changes, and were first explored for assessing circadian rhythms (Witting et al. 1990; Sadeh et al. 2002). However, continual monitoring of locomotor activity was not feasible until recently due to poor battery life, the inability to wirelessly transmit data, and low patient adherence with research-grade instrumentation. Only recently have these technological constraints been overcome. Today, personal activity monitoring devices such as fitness bracelets or patches – also known as "wearables" – are affordable and widely available to public consumers. This is partly due to the global saturation of the smartphone market, which consequently reduced the cost of manufacturing and distributing similar component parts. Today, wearables house sensors that detect heart rate, activity, ambient light, and sleep. These devices have been used in the studies revealing disturbances in 24-hour routine and circadian rhythm associ-



(a) Actigraphy levels (healthy control).

(b) Actigraphy levels (borderline personality disorder).

Figure 2.5: A "double-plot" of wearable accelerometry or actigraphy data demonstrates night-to-night patterns. The x-axis is the date, and the y-axis is time of day. Each day is repeated adjacent to and below the previous day. This aligns the nights of data and can be particularly useful in depicting circadian rhythm sleep disorders. (a) Actigraphy levels in a healthy control. (b) Actigraphy levels in a patient with borderline personality disorder.

ated with neuropsychiatric illnesses such as BD and schizophrenia (Figure 2.5). While only 2-4% of individuals in the United States have a wearable device, the market is estimated to increase to 115 million units in 2018 and generate $50B of revenue (Gandhi et al. 2014; Statista 2017). Here we review recent studies using wearable accelerometers to monitor neuropsychiatric illnesses.

Monitoring schizophrenia via wearable accelerometers

Martin *et al.* found older schizophrenia patients have more disrupted sleep and circadian rhythms (Martin et al. 2005). 28 older schizophrenia patients (mean age=58.3 years) and 28 age- and gender-matched controls were monitored for three days using Actillume wrist actigraphs (Ambulatory Monitoring, Inc., Ardsley, New York). Minute-by-minute activity and light exposure were recorded. Patients spent longer in bed, had more disrupted nighttime sleep, slept more during the day, and had less robust circadian rhythms of activity and light exposure compared to controls.

Apiquian *et al.* evaluated rest-activity characteristics in 20 unmedicated and non-hospitalized schizophrenia patients and 20 controls for five days using a wrist-worn actigraph (Actiwatch-16) (Apiquian et al. 2017). Compared to controls, untreated patients showed significantly lower levels of motor activity and more sleep time.

Walther *et al.* investigated the relationship between objective measures of motor activity

and PANSS scores (Walther et al. 2009b). 55 schizophrenia patients were monitored for 24 hours via wrist actigraphy. Low activity levels were correlated with high PANSS negative syndrome subscale scores. Interestingly, actigraphic parameters did not correlate with motor-specific questions of the PANSS, challenging the validity of the questionnaire.

This same research group subsequently used 24-hour actigraphy to differentiate schizophrenia subtypes in a cohort of 60 hospitalized patients (35 paranoid, 12 catatonic, 13 disorganized) (Walther et al. 2009a). Activity level and movement index (proportion of 2-second periods with nonzero activity) were highest in paranoid schizophrenics, whereas the mean duration of uninterrupted mobility was highest in catatonic schizophrenics.

Berle *et al.* used actigraphy to evaluate patterns of motor activity in 23 schizophrenia patients, 23 depressed patients, and 32 control subjects who did not have a history of mood or psychotic systems (Berle et al. 2010). Total motor activity was lower in patients diagnosed with schizophrenia or depression than in controls. However, IS was 18% higher in schizophrenia patients compared to controls, whereas IS did not differ between depressed patients and controls. IV was 18% lower in schizophrenia patients and 8% lower in depressed patients compared to controls.

Hauge *et al.* revisited this same cohort of patients, but analyzed activity data using Fourier analysis and entropy measurements (Hauge et al. 2011). For each patient, these features were derived from the first 300-minute segment of activity data that contained $\leq$ 4 consecutive minutes of zero activity. RMSSD/SD was significantly lower in schizophrenia patients compared to either depressed patients or controls. Sample entropy of activity was significantly lower in depressed patents compared to either schizophrenia patients or controls. Finally, the ratio between variance of HF power and variance of LF power was significantly higher in depressed patients compared to controls.

Wichniak *et al.* recorded seven days of actigraphy using Actiwatch AW4 devices (Cambridge Neurotechnology Inc., UK) in 73 patients with schizophrenia and 36 age- and sex-matched controls (Wichniak et al. 2011). Mental status was measured via the PANSS and

CDSS questionnaires. Schizophrenia patients had lower mean 24-hour activity and mean 10-hour daytime activity levels, and spent more time in bed. Lower activity was associated with higher PANSS and CDSS scores.

Sano *et al.* recorded seven days of actigraphy (Actigraph Mini-Motionlogger; Ambulatory Monitors Inc., Ardsley, NY, USA) in 19 schizophrenia patients and 11 controls (Sano et al. 2012). Resting periods obeyed a power-law cumulative distribution whereas active periods obeyed a stretched exponential distribution. Distribution parameters differed among schizophrenia patients and controls. For resting periods, the average scaling exponent values (mean $\pm$ standard deviation) were $\bar{\gamma} = 0.86 \pm 0.03$ for schizophrenia patients and $\bar{\gamma} = 0.99 \pm 0.03$ for controls. For active periods, the average stretching parameters were $\bar{\beta} = 0.57 \pm 0.02$ for schizophrenia patients and $\bar{\beta} = 0.64 \pm 0.02$ for controls.

Evaluating the distribution of rest-activity periods was also previously described by Nakamura et al. 2007 and Sano et al. 2012. Fasmer et al. 2015 used this approach in a cohort of 24 patients with schizophrenia, 23 with depression, and 29 controls. 12 days of actigraphy data were recorded per patient using Actiwatches (Cambridge Neurotechnology Ltd., England, UK). For active periods, average scaling exponent values (mean $\pm$ standard deviation) were $\bar{\gamma} = 0.77 \pm 0.13$ for schizophrenia patients, $\bar{\gamma} = 0.88 \pm 0.13$ for depressed patients, and $\bar{\gamma} = 0.82 \pm 0.01$ for controls. For inactive periods, average scaling exponent values (mean $\pm$ standard deviation) were $\bar{\gamma} = 0.81 \pm 0.17$ for schizophrenia patients, $\bar{\gamma} = 0.93 \pm 0.18$ for depressed patients, and $\bar{\gamma} = 0.71 \pm 0.11$ for controls. Length of active and inactive periods and scaling exponents for both active and inactive periods correlated with IS, whereas only length of active periods and scaling exponents for inactive periods correlated with IV. The authors concluded the distribution of active and inactive periods differed in depressed compared to schizophrenic patients.

Shin *et al.* assessed correlations between locomotor activity and symptom severity of 61 subjects with schizophrenia (Shin et al. 2016). Subjects wore a Fitbit Flex device for a week to assess their activity, and completed the PANSS questionnaire to assess schizophrenia

symptoms. Subjects with a high total PANSS score or high positive subscale scores had significantly lower levels of physical activity than the other groups.

2.5.3 Holter monitoring

Much literature has established a bidirectional relationship between changes in HRV and neuropsychiatric illness. People with severe mental illness have worse cardiovascular outcomes than healthy controls, and people with serious cardiovascular illness are more likely to develop certain neuropsychiatric illnesses (Newcomer et al. 2007; Sowden et al. 2009). The interplay between mental and cardiovascular health is believed to be mediated by alterations in the ANS, endocrine effectors such as cortisol and catecholamines, activation of pro-inflammatory cytokines, and lifestyle and environmental exposures such as diet, exercise, and social support (Grippo et al. 2009). In particular, HR and HRV measures can be measured noninvasively, and reflect the state of the ANS. Alterations in HR and HRV may thus provide an objective and passively measurable marker of clinical status in neuropsychiatric illnesses ranging from MDD to BD to schizophrenia (Henry et al. 2010; Cohen et al. 2000).

Cardiac monitoring from body-worn instrumentation is known as Holter monitoring and was originally performed via large stationary ECG devices. Recently, body-worn patches adhering to the skin have been developed to measure HR via ECG, actigraphy, and even metabolites in sweat (Rodgers et al. 2015). Adhesive patches have the potential to improve adherence with study protocols and device use because they are unobtrusive and always attached to the patient. Most studies using physiological patches have focused on heart disease, although a few groups have used this technology to evaluate patients with depression, stress, and schizophrenia. Photoplethysmography (PPG) is another approach for assessing HR via optical measurements of changes in blood volume, and has become a popular sensing technique in wearable devices such as fitness bracelets (Allen 2017). Here we summarize several studies that exclusively focus on the analysis of heart rate data, measured via both traditional ECG as well as patch-based sensing modalities. Devices utilizing PPG are reported

in the next section on multi-modal sensing.

Monitoring schizophrenia via Holter monitor

Cardiovascular mortality risk is elevated in patients with schizophrenia, which may be due to increased prevalence of obesity, smoking, and diabetes, adverse pro-arrhythmic effects of antipsychotic medication, and altered autonomic function. Bär et al. 2017 calculated complexity measures of HRV using short-term ECG recordings from 20 unmedicated subjects with schizophrenia and 20 age- and gender-matched healthy controls. Features included joint symbolic dynamics, compression entropy, fractal dimension and approximate entropy. For analysis of symbolic dynamics, every beat duration was compared to the preceding beat. Whenever the beat durations differed by ≤ 10 milliseconds, the pattern was encoded as '0'. If the beat durations differed by > 10 milliseconds, the pattern was encoded as '1'. A sequence of six consecutive '0's or '1's denoted sequences with low or high variability, respectively. The frequency of low or high variability sequences was used as a feature. Complexity of HR time series was significantly reduced in acute schizophrenia. However, when using HR as a covariate, only fractal dimension remained significantly altered.

d. Overall, the length of two consecutive beats is compared, and a differentiation is being made whether these differ by more or less than the given time limit.

Monitoring PTSD via Holter monitor

Cohen *et al.* evaluated frequency-domain HRV measures via power spectral density analysis using ECG recordings from 14 subjects with post traumatic stress disorder (PTSD), 11 subjects with panic disorder, and 25 matched controls (Cohen et al. 2000). ECG recordings were made while subjects were resting while recalling the trauma implicated in the development of their PTSD, or the circumstances of a severe panic attack, as appropriate, and again while resting. Controls were asked to recall a stressful life event during recall. Both PTSD and panic disorder groups had elevated HR and low frequency LF power at baseline, suggesting

increased sympathetic activity. However, PTSD patients did not respond to recall stress with increases in HR and LF.

Reinertsen *et al.* used a machine learning approach to dichotomize subjects with PTSD from healthy controls using features such as LF power, statistical moments, and acceleration and deceleration capacity (Reinertsen et al. 2017a). 24-hour single-channel ECG recordings were obtained from 23 subjects with current PTSD, and 25 control subjects with no history of PTSD. RR intervals derived from these data were cleaned and used to calculate HR and HRV features – including statistical moments, power spectral density components, entropy, and acceleration and deceleration capacity – which were used to train a logistic regression classifier. Performance was assessed via repeated random sub-sampling validation. To reduce noise and activity-related effects, features were calculated from five non-overlapping ten-minute quiescent segments of RR intervals defined by lowest HR, as well as random ten-minute segments as a control method. Feature selection was performed and a median AUC of 0.86 was achieved out-of-sample test set data. This was significantly higher than the AUC using 24 h of data (0.72) or random segments (0.67), demonstrating the utility of a novel HR segmentation approach for improving the classification of PTSD from HR and HRV measures. Further work should prospectively evaluate if classifier output changes significantly with worsening symptomatology or effective treatment of PTSD.

2.5.4 Multimodal sensing

Here we review studies that utilize heart rate in addition to other sensor types, including accelerometry, ambient light, and GPS. A patient with milder severity of an illness such as schizophrenia may not demonstrate significant alterations in accelerometry or heart rate-derived features in a univariate sense, but multiple weak features can be aggregated together to train a classifier that accurately infers symptomatology or clinical status. However, commercially available devices with physiologically and behaviorally relevant sensing technologies of high accuracy have only recently reached the market. Furthermore, awareness of the

utility of conglomerating several weaker signals is more prevalent amongst machine learning practitioners than statisticians and clinical investigators. Relatively few studies employ multi-sensor fusion approaches, and even fewer focus on neuropsychiatric illness.

Kamdar *et al.* explored the prediction of emotional state from accelerometry, ambient light, and heart rate data measured via Samsung Gear S smartwatches (Kamdar et al. 2016). Data was collected from 13 healthy subjects in a pilot test. A web app was also developed for users to self-report moods via a Likert scale rating of happiness, energy, and relaxation. The app also captured user keystrokes and mouse patterns. Each subject wore the Gear S watch for at least 6 hours and entered at least three self-reported moods over a single day. Several machine learning algorithms were trained using these features: random forest, gradient boosted regressor trees, regularized logistic regression, SVM, and k nearest neighbors. A random forest model explained 51% of the variance of emotional state from device-captured data. However, top features were derived primarily from user interactions rather than passively monitored physiology. Furthermore, no classifier accuracy metrics – such as AUC of classification of mood status – were reported, and the authors also reported high levels of variance in HR measured with the watch compared to a direct pulse measurement, although the latter method was not specified.

AlHanai *et al.* used a combination of auditory, text, and physiological signals to predict the mood (happy or sad) of 31 narrations from ten subjects as they told either happy or sad stories (AlHanai et al. 2017). Subjects wore wrist-mounted Samsung Simband devices which recorded PPG, ECG, accelerometry, skin impedance, galvanic skin response, and skin temperature. Audio was recorded using Apple iPhones. 386 audio and 222 physiological features were calculated from the data. A subset of 4 audio, 1 text, and 5 physiologic features were identified using sequential forward feature selection: subject movement, cardiovascular activity, energy in speech, probability of voicing, and linguistic sentiment (i.e. negative or positive). A deep neural network was trained using these features to classify if the story was happy or sad. To ensure the real-time utility of the model, classification was performed over 5

second intervals. Model performance was assessed via leave-one-subject-out cross-validation, and the classifier achieved a mean AUC of 0.92.

Osipov *et al.* measured HR and accelerometry in 16 subjects with schizophrenia and 19 controls using an adhesive monitoring patch (Protues Digital Health, Redwood City, CA) (Osipov et al. 2015). Features calculated on both types of data included basic summary statistics – mean, median, mode, and variance – as well rest-activity characteristics (Van Someren et al. 1999), multiscale sample entropy (Costa et al. 2002), and multiscale transfer entropy (Schreiber 2000). An SVM learned to dichotomize subjects as either having a diagnosis of schizophrenia or being a control. Two-fold cross-validation with repeated random sub-sampling was performed 1000 times. Using HR features resulted in an AUC of 0.85, whereas using activity features resulted in AUC of 0.90. Using both HR and activity features resulted in an AUC of 0.99.

Reinertsen *et al.* measured HR and locomotor activity in 12 medicated subjects with schizophrenia and 12 healthy controls, and classified contiguous days of data as belonging to a schizophrenia patient or a healthy control (Reinertsen et al. 2017b). Subjects were monitored for 3–4 weeks using a disposable adhesive patch sensor worn on the chest and manufactured by Proteus Biomedical (Redwood City, CA). Features derived from time series data included classical statistical characteristics, rest-activity metrics, transfer entropy, and multiscale $\mathcal{H}_{\text{fuzzy}}$. The analysis window length, or number of days of data considered per record, was varied from two to eight days. An SVM was trained with these features to classify records as belonging to either a schizophrenia or control subject. Model performance was assessed via subject-wise leave-one-out-crossfold-validation. An analysis window length of eight days resulted in a high AUC of 0.96. Reducing the analysis window length to two days only lowered the AUC to 0.91. The type of most predictive features varied with analysis window length. Classifier output may have represented illness severity or level of ANS dysfunction, although verifying this in future work will require gathering information about symptoms on a daily basis.

Cella *et al.* monitored 30 subjects with schizophrenia and 25 controls using wrist-worn Empatica E4 devices which measured skin conductance, PPG (from which RR intervals were derived), and accelerometry (Cella et al. 2017). Symptom severity in subjects with schizophrenia was assessed via the PANSS questionnaire. Subjects were monitored for six days, and recordings < 60 minutes were excluded. At least two 8-hour recordings were obtained for each subject, with an average of 3-4 8-hour recordings obtained per subject. Skin conductance did not vary by patient group, but subjects with schizophrenia had significantly lower SDNN and RMSSD values, as well as lower locomotor activity and fewer hours of structured activity, compared to controls. Chlorpromazine levels were not found to significantly affect any physiological measures.

2.6 Conclusion

In closing, many studies have explored the use of smartphones, wearable accelerometers, Holter devices, and multimodal sensors for monitoring of patient physiology, psychology, and behavior. These technologies continue to decrease in cost and permeate other facets of daily life. However, several technical challenges remain that if addressed could improve the accuracy, interpretability, and potential usability of passive monitoring for research and patient care. Most analyses are performed on all available data in an effort to maximize the amount of information that may contain clinically relevant insights, despite the presence of noise and endogenous behavioral or physiological signals unrelated to illness throughout a recording. The relationship between classifier accuracy and quality of EKG data has been evaluated, but the influence of data quantity, time of recording, and other contextual information on the performance of a machine learning algorithm must still be studied to better inform monitoring approaches. Finally, univariate measures of signal regularity or complexity over multiple time scales have been found to correlate with illness, but the dynamics of how different physiological and behavioral signals interact in an information-theoretical sense, and the association of these dynamics with disease, remains unexplored.

CHAPTER 3
CLASSIFICATION OF PTSD FROM HEART RATE DATA

3.1 Overview

Objective. HRV characterizes changes in autonomic nervous system function and has been shown to vary with post traumatic stress disorder (PTSD). However, HR data has noise and artifactsdue to intrinsic issues with measurement as well as volitional movement and behavior.

Approach. We proposed to improve the signal-to-noise ratio of HR data via a HR-based window segmentation approach whereby five 10-minute segments with the lowest median HR are isolated. These segments may represent quiescent periods, or times when the subject was least likely to be engaged in volitional motor activity and/or most likely to be sleeping or resting. To validate the approach, single-channel ECG data were collected from 23 subjects with current PTSD, and 25 control subjects with no history of PTSD over 24 hours. RR intervals were derived from these data, cleaned, and used to calculate HR and HRV metrics. Features were derived from 1) RR data from these segments, 2) RR data from five randomly selected 10-minute control segments, or 3) all 24 hours of RR data. Classifier performance was assessed via repeated random sub-sampling validation, and area under the receiver operating characteristic curve (AUC) was calculated.

Main results. A combination of the four most predictive features derived from quiescent segments resulted in a median area under the receiver operating curve (AUC) of 0.86 on out-of-sample test set data. This was significantly higher than the AUC using 24 hours of data (0.72) or random segments (0.67).

Significance. These results demonstrate our segmentation approach improves the classification of PTSD from HR and HRV measures, and suggest the potential for tracking PTSD illness severity via objective physiological monitoring (Reinertsen et al. 2017a). Future stud-

ies should prospectively evaluate if classifier output changes significantly with worsening or effective treatment of PTSD. Determining if this approach is useful for clinical decision support will require a larger randomized controlled trial with monitoring connected to specific outcomes.

3.2 Motivation and study organization

Patients with PTSD have significantly different measures of heart rate variability (HRV) compared to healthy controls (Liddell et al. 2016; Minassian et al. 2014). HRV – changes in beat-to-beat heart rate – can be used to assess changes in the autonomic nervous system (Clifford 2002; Pan et al. 2016). Recently, twins with PTSD were reported to have 49% lower low frequency (LF) HRV compared to their brothers without PTSD (Shah et al. 2013). When attempting to identify differences in autonomic function as measured by HRV, it is important to control for other factors such as stress, affect, physical activity, and cardiovascular or neurological disease other than PTSD.

Evaluating HRV during sleep can account for confounding from stress, affect, and physical activity (Germain et al. 2005). Furthermore, some reports show HRV reductions due to PTSD are greatest during the night (Woodward et al. 2009; Kobayashi et al. 2014), suggesting that analyzing data only during nocturnal sleep could improve classifier performance. However, HRV metrics vary by sleep stage due to changes in vagal and sympathetic activity during REM, light and deep sleep (Vanoli et al. 1995; Viola et al. 2002). Segmentation by sleep stage may thus improve the signal to noise ratio. For example, in earlier work using this novel methodology, we showed that comparing HRV metrics in REM sleep, and separately in deep sleep, better separated sleep apneic patients from healthy controls (Clifford et al. 2004). This approach may also apply to other illnesses associated with changes in HRV measures, such as PTSD. However, accurately measuring sleep status or estimating sleep stage from other data such as HR is difficult.

PTSD has been classified using self-reported data and demographics (Kessler et al. 2014;

Karstoft et al. 2015; Galatzer-Levy et al. 2014). However, a multivariate classifier separating PTSD patients and controls using HRV measures or other objective physiological data has not yet been developed. Additionally, the utility of thresholding on individual HRV measures to identify PTSD has yet to be evaluated.

Here we propose a novel method of controlling for activity by only evaluating quiescent segments of RR intervals, with quiescence determined by lowest median HR for each subject. This segmentation approach may reduce random error from mental and physical activity, highlight involvement of the autonomic nervous system, and approximate restfulness in the absence of validated sleep stage data.

The objectives of this work are to: 1) calculate features from HR and HRV measures indicative of PTSD in male veterans using 24-hour Holter ECG recordings, 2) use these features to train a multivariate classifier whose output – a probability of membership in either the PTSD or control group – could potentially be used as a proxy for illness severity in a patient already diagnosed with PTSD, and 3) improve classifier performance using a novel segmentation method on RR intervals to reduce noise and potential confounders.

All data processing, feature extraction, and classifier training was performed using Matlab R2016b (Mathworks, Natick, MA).

3.3 Methods

3.3.1 Subject enrollment

ECG recordings were obtained from 24 male subjects with current PTSD (symptoms within the past 30 days) and 26 healthy control subjects in a dataset derived from the Emory Twins Studies first reported by Shah et al. 2013. This smaller cohort was selected to balance classes, i.e. a similar number of subjects with PTSD as controls. Participants were subjects with clinical diagnoses of PTSD, and healthy control subjects examined at the same time at the Emory University General Clinical Research Center. Individuals lacking sufficient ECG data were excluded (see exclusion criteria in later sub-section). All participants wore

an ambulatory ECG (Holter) monitor (GE Marquette SEER digital system; GE Medical Systems, Waukesha, WI) for 24 hours. Participants had matched recording times and schedules. Activity was restricted to non-strenuous walking around the university campus and medical center, and participants were told to refrain from smoking or drinking alcohol or coffee. This study was approved by the Emory Institutional Review Board (81004), and all subjects signed an informed consent.

3.3.2 Data recording

The ECG signal was sampled at 125 Hz. Data were downloaded to a local HIPAA-compliant data repository using a MARS SEER Light digital recorder. QRS complexes were detected and annotated in the ECG automatically using the GE MARS software. RR intervals were calculated from the time difference between adjacent annotated beats.

3.3.3 Data pre-processing and exclusion criteria

RR intervals obtained later than 24 hours after the start of recording were discarded. Ectopic beats and artifacts were removed via established methods (Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology 1996); non-physiological RR intervals with values >1.5 seconds or <0.33 seconds were discarded, and RR intervals 20% shorter or longer than the previous RR interval, or 20% shorter or longer than the overall mean RR interval were discarded. Gaps in the time series were interpolated via linear spline. RR intervals were re-sampled at 3.413 Hz (1024 samples per five minute segment) to create a uniformly spaced time series for spectral HRV measures. One subject with PTSD and one subject without PTSD had fewer than 22 hours of ECG recordings; both were excluded from further analysis. Cleaned RR intervals were obtained from 23 subjects with PTSD and 25 control subjects (48 total). To demonstrate the utility of data pre-processing, uncleaned RR intervals were also evaluated as a comparison.

3.3.4 Identification of quiescent segments

To reduce confounding effects of mental and physical activity, five non-overlapping ten-minute periods with the lowest median HR – hereafter referred to as "quiescent segments" – were identified from cleaned RR data for each subject. Figure 3.1illustrates a representative 24-hour RR tachogram from a study subject, with quiescent segments indicated by shaded regions. Healthy humans cycle through each of the five defined sleep stages with a period of approximately 100 minutes, and each sleep stage lasts up to 20 minutes; this informed our selection of segment length (Clifford et al. 2004). For each subject, each feature was calculated for each of five quiescent segments, resulting in $5 \times m$ total features per subject. For each feature, the median feature value from the five segments was calculated, resulting in m features per subject to be used for training a logistic regression model. Feature extraction was also performed on ten-minute segments chosen at random, excluding quiescent segments of lowest HR, to serve as a control and to investigate if segment length was a confounder.

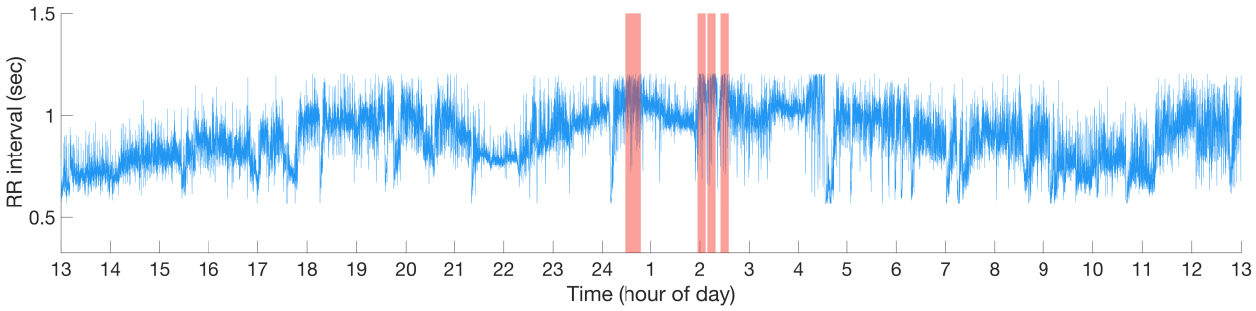


Figure 3.1: Representative time series of RR interval data from a single subject. Shaded red areas are ten-minute quiescent segments. Horizontal axis is time of day in hours; 13 corresponds to 1 PM, 1 corresponds to 1 AM, etc. ECG recording started at the origin of the x-axis (approximately 1 PM).

3.3.5 Feature extraction and Heart Rate Variability measures

Cleaned RR intervals from either a) all 24 hours of ECG recordings, b) random control segments, or c) quiescent segments were used to calculate features. These features included the median quiescent window time converted to radians, basic RR interval statistics (mean,

median, mode, standard deviation (σ_{rr}), interquartile range (IQR_{rr}), skewness, and kurtosis), AC, DC, power spectral measures (VLF, LF, HF, total power), and other measures of the distribution of RR intervals (NNN, MNN, PNN, PNN50, RMSSD, and SDNN) (Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology 1996).

3.3.6 Power spectral measures of HRV

HRV power spectral measures were computed from cleaned RR interval time series with a fast Fourier transform (FFT) and a Parzen window, following our previous methodology (Shah et al. 2013). The FFT and spectra were corrected for window attenuation and boxcar sampling. The power spectrum was integrated over four discrete frequency bands: ultra-low frequency (ULF) <0.0033 Hz; very low frequency (VLF) 0.0033 – 0.04 Hz; low frequency (LF) 0.04 – 0.15 Hz; and high frequency (HF) 0.15 – 0.40 Hz (Bigger et al. 1992). These frequency bands measure the renin-angiotensin, sympathetic, and parasympathetic cardiovascular control systems (Akselrod et al. 1981). Total power, incorporating the full spectrum from 0 – 0.40 Hz was also estimated.

3.3.7 Phase-rectified signal averaging

Phase-rectified signal averaging (PRSA) was performed on cleaned RR intervals to quantify acceleration and deceleration capacity of HR. This method can be used to detect quasi-periodic oscillations and to separate processes occurring during increasing and decreasing parts of the signal (Bauer et al. 2006b). Furthermore, PRSA is robust to noise and non-stationarity. Heartbeat interval shortenings are used as anchors for acceleration-related PRSA signals, whereas heartbeat interval lengthenings are used as anchors for deceleration-related PRSA signals. Sampling frequency was set to 512 Hz, and the window length around anchors was set to 30 elements.

3.3.8 Assessment of PTSD

The Structured Clinical Interview for Psychiatry Disorders was administered to classify subjects into two classes: 1) current PTSD with symptoms within the past 30 days, or 2) no history of PTSD (control subjects).

3.3.9 Feature selection and classification

All twenty features, as well one feature at a time, were used to train a logistic regression. The output of this binary classifier was the probability of membership in the PTSD class. L1L2 (elastic net) regularization was performed to reduce coefficient values for collinear or non-predictive features and create a sparser and more generalizable model. Unconstrained differentiable multivariate optimization was performed using `minFunc`. Specifically, maximum likelihood estimation was performed via quasi-Newton limited-memory Broyden-Fletcher-Goldfarb-Shanno updating (Bishop 1995). Distributions of features from PTSD and control subjects were visualized and compared via two-sided Kolmogorov-Smirnov tests. Additionally, given the relatively low number of features, a grid search was performed to assess combinations of features.

To assess classifier performance on out-of-sample data, we performed bagging with replacement, an ensemble method to reduce variance and avoid overfitting (Breiman 1996; Arlot et al. 2010). Data were randomly split into training and test data at a 70:30 ratio, with the class balance in training and test sets maintained to reflect the class balance in the entire data set. By random sampling with replacement, some data may be used more than once between models, or not be used at all. Features in the training set were transformed to have Gaussian distributions using either the identity, square root, or logarithmic transformations. The transformation which provided the lowest k-statistic using the Lilliefors test was used on both training and test sets. Data were then z-scored to by subtracting the training mean and dividing by the training standard deviation on both the training and test sets. A grid search was performed to select the value of λ ranging from 0.001 – 5.0 that maximized

the test set AUC within the model. The classifier thus learned solely from training data, and was evaluated solely on test data. Sampling, feature transformation, learning, and classifier evaluation was repeated nine more times for a total of ten models. AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for training and test sets within each model.

3.4 Results

3.4.1 Temporal distribution of quiescent segments

The temporal distribution of quiescent segments does not differ by PTSD status ($P = 0.23$ via two-sided Kolmogorov-Smirnov test; Figure 3.2). Box plots are not associated with the y-axis; + indicates the mean, the middle line indicates the median, the box denotes the interquartile range (IQR) flanked by the 25th and 75th percentiles, the vertical lines outside of the box indicate the 9th and 91st percentiles, and circles indicate outliers.

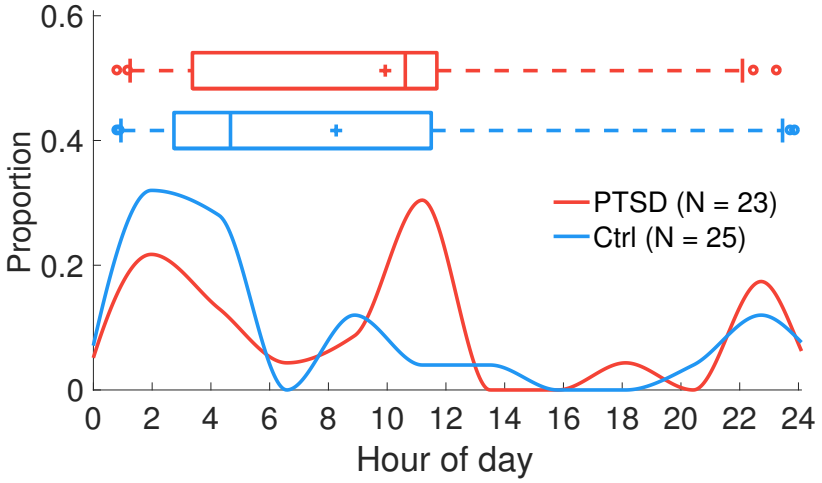


Figure 3.2: Temporal distribution of quiescent segments does not differ by PTSD status ($P = 0.23$). The x-axis denotes hour of the day (i.e. hours past midnight), ranging from 0 to 24; 12 corresponds to noon. Red indicates quiescent segments from subjects with PTSD (23 subjects); blue indicates quiescent segments for healthy controls (25 subjects).

3.4.2 Classifier trained on all features

All twenty features were used to train an L1L2 regularized logistic regression. Classifier performance was assessed for three different segmentation approaches, and using either uncleaned or cleaned RR intervals. Using quiescent segments of cleaned RR intervals results in greater predictivity compared to other segmentation approaches, with a training AUC of 0.87 and a test AUC of 0.70 (Table 3.1).

Table 3.1: AUCs of L1L2 regularized logistic regression models using all HR and HRV features extracted from RR intervals. Values shown are medians and IQR bounds in brackets.

	Train AUC		Test AUC	
	No RR cleaning	RR cleaning	No RR cleaning	RR cleaning
24 hours	0.77 [0.75 0.82]	0.75 [0.70 0.78]	0.54 [0.46 0.64]	0.58 [0.46 0.64]
Random segments	0.76 [0.73 0.80]	0.78 [0.77 0.80]	0.50 [0.45 0.57]	0.56 [0.50 0.71]
Quiescent segments	0.89 [0.87 0.91]	0.87 [0.83 0.89]	0.73 [0.70 0.80]	0.75 [0.71 0.82]

3.4.3 Classifier trained on individual features and combinations of features

To improve classifier performance, individual features and combinations of features were used to train a regularized logistic regression. Testing many combinations of features is computationally inefficient, but was feasible here given the small number of features and fast speed of training a logistic regression model. Classifier performance was assessed for three different segmentation approaches, using uncleaned or cleaned RR intervals. A classifier trained on the most predictive combination of four features derived from quiescent segments of RR intervals achieves greater predictivity (training AUC = 0.85, test AUC = 0.84) compared to when using features derived from random segments or 24 hours of RR intervals (Table 3.2).

The most predictive combination of four features derived from 24 hours of RR intervals is σ_{rr}, $\mathrm{IQR}_{\mathrm{rr}}$, LF power, and SDNN (Table 3.3). The most predictive combination of four features derived from quiescent segments of RR intervals were AC, DC, LF power, and SDNN (Table 3.4). The β coefficients of these most predictive models are shown in Table 3.5, and

other classifier performance metrics are shown in Table 3.6.

Table 3.2: AUCs of L1L2 regularized logistic regression models using the top four features extracted from RR intervals. Values shown are medians across sub-samples and IQR bounds in brackets.

	Train AUC		Test AUC	
	No RR cleaning	RR cleaning	No RR cleaning	RR cleaning
24 hours	0.74 [0.73 0.78]	0.73 [0.69 0.74]	0.66 [0.45 0.66]	0.67 [0.62 0.71]
Random segments	0.70 [0.66 0.76]	0.76 [0.72 0.77]	0.61 [0.50 0.64]	0.72 [0.62 0.77]
Quiescent segments	0.85 [0.84 0.88]	0.85 [0.83 0.88]	0.81 [0.70 0.84]	0.86 [0.75 0.88]

Table 3.3: Features extracted from 24 hours of of RR intervals, shown as medians and IQR bounds in brackets. CTRL refers to the control group. Test AUC reports performance of univariate classifier trained solely on one feature.

	PTSD status		
Feature	PTSD	CTRL	Test AUC
AC (sec)	−8.28 [−1.27e1 −6.31]	−1.04e1 [−1.33e1 −8.18]	0.54 [0.52 0.68]
DC (sec)	8.19 [6.55 1.23e1]	1.05e1 [8.89 1.38e1]	0.58 [0.54 0.73]
LF power (sec^2)†,*	3.51e2 [1.37e2 4.91e2]	5.86e2 [3.76e2 8.76e2]	0.71 [0.64 0.80]
σ_{rr} (sec)*	1.15e−1 [9.15e−2 1.34e−1]	1.29e−1 [1.14e−1 1.51e−1]	0.65 [0.59 0.73]
IQR_{rr} (sec)*	1.76e−1 [1.26e−1 2.11e−1]	2.08e−1 [1.52e−1 2.34e−1]	0.63 [0.59 0.67]
SDNN (sec)*	3.89e1 [2.97e1 5.42e1]	5.07e1 [4.09e1 6.32e1]	0.61 [0.55 0.75]

†: $P < 0.05$ comparing feature values from PTSD vs. control subjects via two-sided Kolmogorov—Smirnov test.
⋆: Feature among combination that maximizes training set AUC.

3.4.4 Distributions of predictive features

Distributions of predictive features were visualized (Figure 3.3 – Figure 3.5). Box plots are not associated with the y-axis; + indicates the mean, the middle line indicates the median, the box denotes the IQR flanked by the 25th and 75th percentiles, the vertical lines outside of the box indicate the 9th and 91st percentiles, and circles indicate outliers.

Segmentation improves separability of some features as determined by two-sided Kolmogorov-Smirnov tests. AC does not significantly differ by PTSD status when evaluating 24 hours of data ($P = 0.24$), but is significantly higher in subjects with PTSD versus controls when analyzing quiescent segments ($P = 0.04$; Figure 3.3). Similarly, DC does not significantly

Table 3.4: Features extracted from quiescent segments of of RR intervals, shown as medians and IQR bounds in brackets. CTRL refers to the control group. Test AUC reports performance of univariate classifier trained solely on one feature.

	PTSD status		
Feature	PTSD	CTRL	Test AUC
AC (sec)†,*	−9.62 [−1.26e1 −6.22]	−1.28e1 [−1.91e1 −9.72]	0.77 [0.73 0.82]
DC (sec)†,*	9.43 [6.64 1.22e1]	1.40e1 [1.11e1 2.06e1]	0.82 [0.73 0.84]
LF power (sec^2)†,*	3.31e2 [1.52e2 5.78e2]	8.71e2 [4.44e2 1.47e3]	0.81 [0.75 0.88]
σ_{rr} (sec)†	4.14e−2 [3.44e−2 5.34e−2]	7.12e−2 [4.9e−2 8.06e−2]	0.82 [0.73 0.84]
IQR_{rr} (sec)†	5.40e−2 [3.55e−2 5.60e−2]	7.20e−2 [5.50e−2 9.38e−2]	0.78 [0.71 0.81]
SDNN (sec)†,*	4.68e1 [3.16e1 5.97e1]	6.47e1 [4.32e1 7.70e1]	0.75 [0.57 0.86]

†: $P < 0.05$ comparing feature values from PTSD vs. control subjects via two-sided Kolmogorov—Smirnov test.

⋆: Feature among combination that maximizes training set AUC.

Table 3.5: β coefficients of L1L2 regularized logistic regression models trained on four most predictive features from either 24 hours or quiescent segments of RR intervals. Values shown are medians across sub-samples and IQR bounds in brackets.

	24 hours		Quiescent segments	
	Feature	Coefficient value	Feature	Coefficient
β_1	Intercept	0.06 [0.05 0.06]	Intercept	0.08 [0.08 0.11]
β_2	σ_{rr}	0.46 [0.35 0.51]	AC	1.12 [1.03 1.60]
β_3	IQR_{rr}	0.29 [0.22 0.54]	DC	0.80 [0.61 1.06]
β_4	LF power	0.00 [-0.03 0.07]	LF power	0.32 [0.00 0.67]
β_5	SDNN	-0.04 [-0.31 -0.00]	SDNN	0.30 [0.01 0.39]

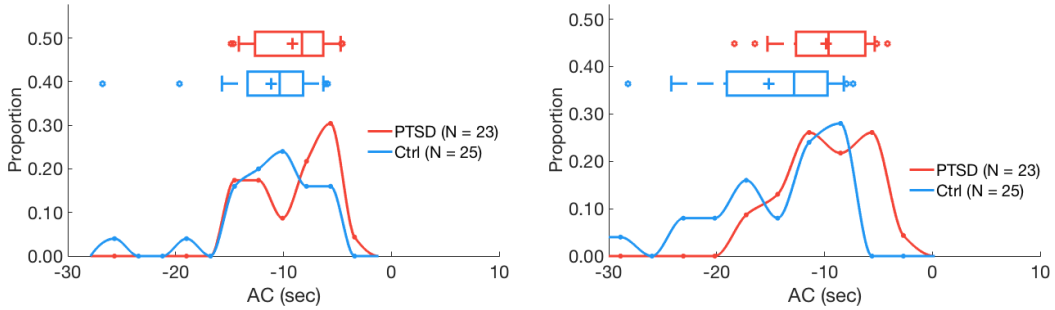
differ by PTSD status when evaluating 24 hours of data ($P = 0.13$), but is significantly lower in subjects with PTSD versus controls when analyzing quiescent segments ($P = 0.01$; Figure 3.4). LF power is lower in PTSD for both 24-hour data ($P = 0.01$) and quiescent segments of data ($P = 0.01$; Figure 3.5). SDNN does not differ by PTSD status for 24 hours of data ($P = 0.06$), but is significantly lower in PTSD when analyzing quiescent segments ($P = 0.04$; Figure 3.8).

3.5 Discussion

In this study on 23 subjects with current PTSD and 25 controls, HR and HRV features were calculated and used to train an L1L2 regularized logistic regression to classify PTSD status.

Table 3.6: Classifier performance on test set data using most predictive logistic regression models trained on features extracted from RR intervals after using three different segmentation approaches. Values shown are medians across sub-samples and IQR bounds in brackets. PPV is positive predictive value and NPV is negative predictive value.

	Segmentation approach		
Metric	24 hours	Random segments	Quiescent segments
AUC	0.67 [0.62 0.71]	0.70 [0.62 0.79]	0.86 [0.75 0.88]
Accuracy	0.73 [0.67 0.73]	0.73 [0.67 0.80]	0.80 [0.73 0.80]
Sensitivity	0.57 [0.43 0.71]	0.43 [0.43 0.57]	0.71 [0.57 1.00]
Specificity	0.94 [0.75 1.00]	1.00 [0.88 1.00]	0.94 [0.88 1.00]
PPV	0.92 [0.71 1.00]	1.00 [0.78 1.00]	0.94 [0.83 1.00]
NPV	0.69 [0.67 0.75]	0.67 [0.64 0.73]	0.79 [0.73 0.88]



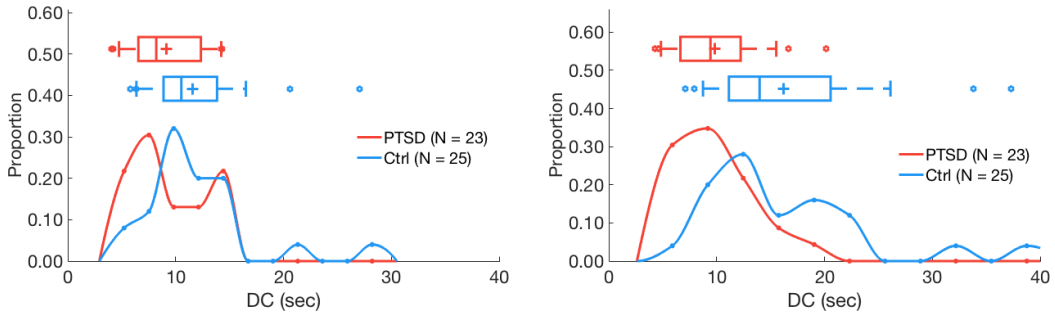
(a) 24 hours of data

(b) Quiescent segments

Figure 3.3: Acceleration capacity (AC) does not differ by PTSD status for 24 hours of RR intervals (a; $P = 0.18$) but is higher in subjects with PTSD for quiescent segments (b; $P < 0.05$).

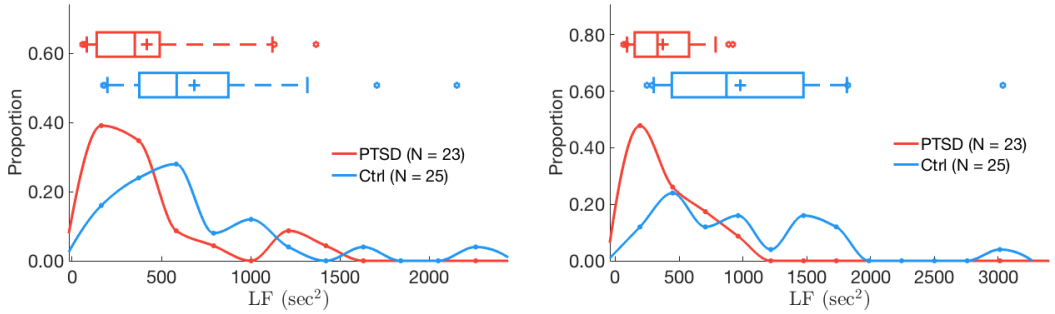
A classifier trained on a combination of the four most predictive features – LF power, σ_{rr}, IQR_{rr}, and SDNN for 24 hours of RR intervals, and AC, DC, LF power, and SDNN for quiescent segments – achieved out-of-sample test AUCs of 0.67 using 24 hours of RR interval data, 0.72 using random segments, and 0.86 using quiescent segments. The simple HR-based window segmentation approach isolated data with the highest signal-to-noise by definition, since data with "signal" was implicitly defined as that with information that maximally improved predictivity of the classifier.

Sleep disordered breathing and sleep disruption are both associated with PTSD, so proxies of sleep are expected to differ by PTSD status (Yesavage et al. 2014; Germain 2013).



(a) 24 hours of data (b) Quiescent segments

Figure 3.4: Deceleration capacity (DC) does not differ by PTSD status for 24 hours of RR intervals (a; $P = 0.09$) but is lower in subjects with PTSD for quiescent segments (b; $P < 0.05$).

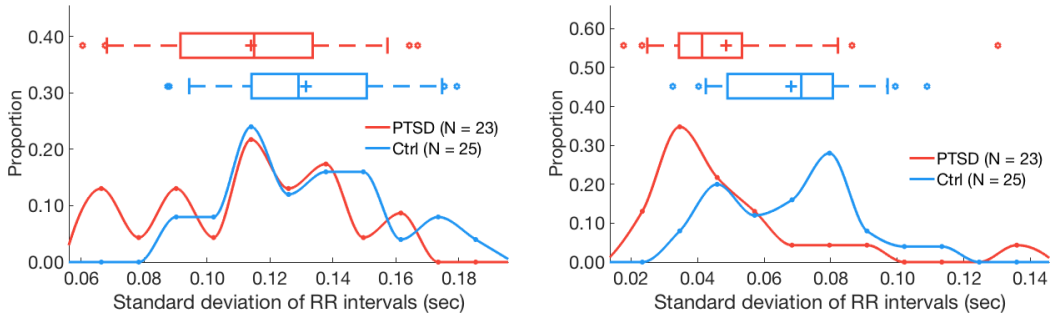


(a) 24 hours of data (b) Quiescent segments

Figure 3.5: Low frequency (LF) power differs by PTSD status for both 24 hours of RR intervals (a; $P < 0.05$) and quiescent segments (b; $P < 0.05$).

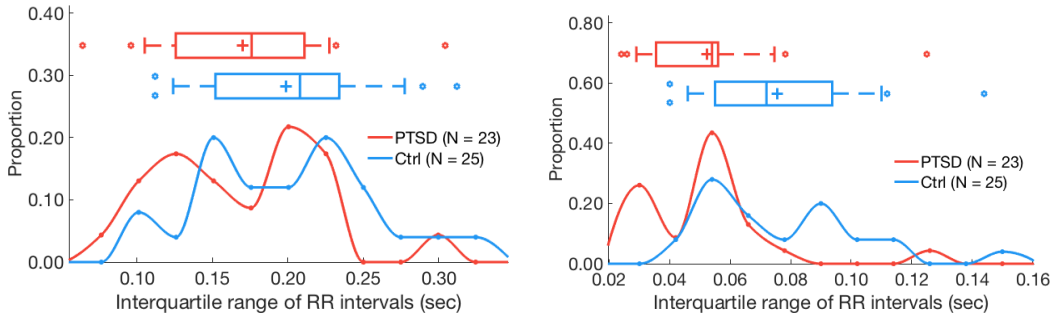
However, the time of median quiescent segments did not significantly differ with PTSD status ($P = 0.23$; Figure 3.2), indicating these factors were not significant in this cohort. Most quiescent segments occurred from midnight to early morning in control subjects. A larger portion of segments were distributed closer to noon in subjects with PTSD. Periods of low HR – a measure of restfulness, not sleep stage – can occur at any time and may reflect differences in sleep patterns, differences in activity, or both. Quiescent segments may contain less noise and movement artifact, as well as reflect lower levels of mental and physical activity, and thus improve the performance of a classifier trained on features from those segments.

Next, all HR and HRV measures were used as features for a logistic regression classifier. L1L2 regularization was performed to reduce coefficient values associated with collinear or



(a) 24 hours of data (b) Quiescent segments

Figure 3.6: σ_{rr} (standard deviation of RR intervals) does not differ by PTSD status for 24 hours of RR intervals (a; $P = 0.25$) but but is higher in control subjects for quiescent segments (b; $P < 0.05$).

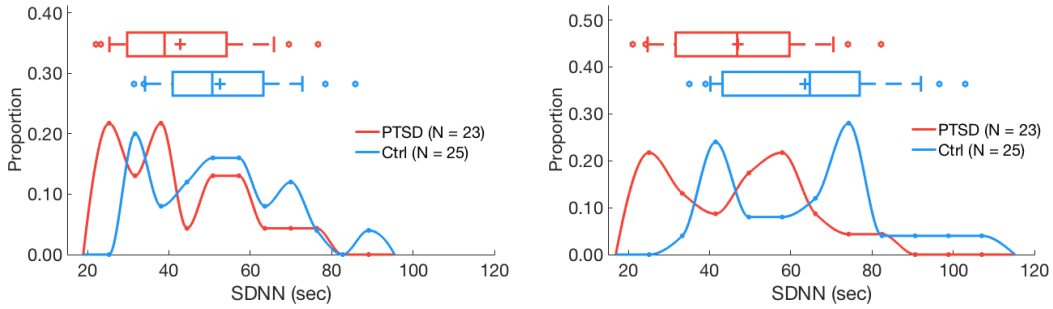


(a) 24 hours of data (b) Quiescent segments

Figure 3.7: IQR_{rr} (interquartile range of RR intervals) does not differ by PTSD status for 24 hours of RR intervals (a; $P = 0.47$) but is higher in control subjects for quiescent segments (b; $P < 0.05$).

redundant features, and to isolate predictive features. A classifier trained on all 20 features from 24 hours of RR intervals achieved a low test AUC of 0.58 (Table 3.1). Using features extracted from quiescent segments improved the test AUC to 0.75, whereas the use of randomly selected control segments resulted in a low test AUC of 0.56. Compared to these low test AUCs, training AUCs were 0.75, 0.78, and 0.87 for 24 hours, random segments, and quiescent segments of RR intervals respectively. These results show a model using all features over-fits training data and would not generalize to out-of-sample data despite regularization. Classifier performance was similar when using uncleaned RR interval data.

Regularization attempts to reduce co-linearity by effectively placing a prior on model coefficients, forcing sparsity with small weights. However, the posterior – formed by updating



(a) 24 hours of data (b) Quiescent segments

Figure 3.8: Standard deviation of normal-to-normal RR intervals (SDNN) does not differ by PTSD status for 24 hours of RR intervals (a; $P = 0.06$) but is higher in control subjects for quiescent segments (b; $P < 0.05$).

Table 3.7: Standard logistic regression on all subjects (N=48) using predictive features extracted from 24 hours of cleaned RR intervals (RR_i). OR is odds ratio, and CI is confidence interval.

Feature	β coefficient	P-value	OR [95% CI]
σ_{rr}	0.08	0.90	1.1 [3.1e-01 3.8]
IQR_{rr}	0.10	0.84	1.1 [4.3e-01 2.9]
LF	-0.36	0.44	7.0e-01 [2.8e-01 1.7]
SDNN	-0.01	0.99	9.9e-01 [3.3e-01 2.9]

the prior with evidence – determines the final form of a model. Thus, with small data sets, even regularized models trained with many features may not work well compared to the use of a hard prior via manual feature selection. Therefore, individual features and combinations of features were used to train lower-dimensional models.

Given $m = 20$ total features and a subset of $k = 1, 2, \ldots, m$ features, the number of possible combinations (i.e. possible arrangements of k features) is the binomial coefficient $\binom{m}{k}$. To ensure feasible computation time and a parsimonious and interpretable model, the maximum number of features used in a combination was limited to four, i.e. $k = 1, 2, \ldots, 4$. Furthermore, using more than four features led to the selection of colinear features and overfitting on the training data (results not shown).

Values of some individually predictive features, and test set AUC and accuracy for classifiers trained these features, are shown in Table 3.3. Distributions of some features were

Table 3.8: AUCs of L1L2 regularized logistic regression models using the top four features extracted from cleaned RR intervals, for either all subjects (N=48) or just paired twins (N=38). Values shown are medians across sub-samples and IQR bounds in brackets.

	All subjects		Paired twins	
	Train AUC	Test AUC	Train AUC	Test AUC
24 hours	0.73 [0.69 0.74]	0.67 [0.62 0.71]	0.70 [0.66 0.72]	0.64 [0.56 0.75]
Random segments	0.74 [0.73 0.80]	0.74 [0.57 0.77]	0.79 [0.75 0.83]	0.76 [0.58 0.78]
Quiescent segments	0.85 [0.83 0.88]	0.86 [0.75 0.88]	0.86 [0.82 0.88]	0.81 [0.69 0.86]

compared via a two-sided Kolmogorov-Smirnov test, but selected the most predictive comnbination features on the basis of maximizing training AUC. For classification, features should be chosen on the basis of predictability rather than significance, because significance alone does not guarantee predictability (Lo et al. 2015). For 24 hours of RR intervals, LF power significantly differed by PTSD status and was one of the four most predictive combination of features (Table 3.3). The other most predictive features were σ_{rr}, IQR_{rr}, and SDNN, but these did not significantly differ by PTSD status. For quiescent segments, the median value of the four most predictive combination of features – AC, DC, LF power, and SDNN – significantly differed by PTSD status (Table 3.4).

AC did not differ by PTSD status for 24 hours of RR intervals, but was higher in subjects with PTSD for quiescent segments (Figure 3.3). Similarly, DC did not differ by PTSD status for 24 hours of RR intervals, but was lower in subjects with PTSD for quiescent segments (Figure 3.4). AC may reflect physiologic performance when parasympathetic withdrawal occurs, whereas DC measures general parasympathetic augmentation (Bauer et al. 2006b; Pan et al. 2016). Although some literature suggests that AC also measures sympathetic activation, this is unlikely because sympathetic modulations occur at 0.1 Hz, which may be four times faster than the modulation frequency of AC, depending on the underlying heart rate (Julien 2006).

LF power differed by PTSD status for both 24 hours and quiescent segments of RR intervals (Figure 3.5). Differences in these measures by PTSD status may be exacerbated in

quiescent segments. In PTSD, vagal augmentation is expected during slow wave sleep, which may be altered by increased insomnia or sleep-disordered breathing. Other physiologic pathways may also be affected during abnormal sleep episodes; low LF may reflect baroreceptor insensitivity (Khoury et al. 2012). These findings underscore physiologic changes that occur with PTSD.

When shifting from 24 hours to quiescent segments, σ_{rr} and $\mathrm{IQR}_{\mathrm{rr}}$ became less predictive, whereas AC and DC became more predictive. In quiescent segments, σ_{rr} in controls was greater than σ_{rr} in subjects with PTSD (Figure 3.6). σ_{rr}, $\mathrm{IQR}_{\mathrm{rr}}$, and SDNN measure variability of RR intervals, and were all significantly lower in quiescent segments from subjects with PTSD. This finding is consistent with previous reports of lower variability of HR being associated with PTSD (Tan et al. 2009; Tan et al. 2011). Additionally, the lack of significance or predictivity of these features (aside from SDNN, which was a predictive feature) from 24 hours of RR intervals is unsurprising because quiescent segments were selected on the basis of low resting HR values, which excludes periods with higher variability. Concerning AC and DC, quiescent segments approximate restfulness rather than sleep state, but may also correspond to slow-wave sleep, during which vagal activity may be augmented and the predictivity of PRSA measures increased.

We calculated β coefficients of L1L2 regularized logistic regressions trained on four most predictive combination of features from 24 hours or quiescent segments of RR intervals (Table 3.5). Although LF power and SDNN were among the most predictive features when using either 24 hours or quiescent segments of RR intervals, the β coefficients of these features significantly differed depending on the segmentation approach. For example, the median β coefficient for LF power computed from 24 hours of RR intervals was close to zero, but for quiescent segments, the median β coefficient was 0.32. This difference suggests interactions between features that reflect the complexity of the underlying physiology, and/or a dependence on time scale.

A regularized classifier trained on the most predictive combination of four features from

a) quiescent segments outperformed classifiers using b) all 24 hours of RR intervals, or c) on random control segments, with test AUCs of a) 0.86, b) 0.67, and c) 0.70 respectively (Table 3.6). Using quiescent segments instead of 24 hours of RR intervals improved every performance metric except specificity, which did not change. Using quiescent segments instead of random segments improved every performance metric except specificity and PPV, which decreased. This suggests classifier performance depends on the information within segments rather than the quantity of data.

We also compared the distribution of classifier output using a Wilcoxon signed rank test to account for the paired nature of these data, and found a statistically significant difference between $P_{estimated}$(PTSD | features from subjects with PTSD) and $P_{estimated}$(PTSD | features from control subjects) ($P<0.001$). This suggests the classifier accurately discriminated PTSD status.

Here the AUC can be interpreted as the ability of a model to classify PTSD status using disease-associated physiological changes. Although learning was done with data from healthy controls, this approach would be suited for monitoring patients with established PTSD. It would not be a screening test for the general population. Future studies could assess how treatments affect physiology, and classify or even predict post-intervention recovery.

We note several limitations of our study. First, our cohort consisted only of 23 subjects with PTSD and 25 controls. This small sample size may not have been adequately powered to detect smaller effect sizes. Our study design would be more elegant with discordant pairs only; however, this would eliminate ten unpaired twins and could reduce statistical power. To evaluate this we compared classifier performance using all subjects (N=48) versus using only paired twins (N=38) (Table 3.8). No statistically significant differences were found using a two-sided Wilcoxon rank-sum test between all subjects and only paired twins cohorts in training or test AUCs for any segmentation approach. This may be due to two competing effects. Reducing sample size could diminish the ability of the classifier to learn predictive features, and decrease out-of-sample test set performance by learning features

not representative of the population distribution. Furthermore, HRV may be about 50% heritable (Su et al. 2010). If HRV and PTSD share an underlying genetic and physiological cause, and our approach evaluates features related to this mechanism, adding paired twins could confound the study, enrich both positive and negative classes with similar physiology-based features, and reduce classifier performance. However, focusing on twins could reduce the random error caused by differences in cardiovascular or autonomic physiology between subjects. Our results suggest the inclusion of non-twins does not reduce the impact of our findings, since we aimed to developed a system for monitoring physiology of subjects with PTSD rather than for screening a correlated population.

A second limitation of our work was only recording 24 hours of ECG data per subject. Our approach could potentially enable home-based continuous physiologic monitoring of the efficacy of a PTSD intervention. However, doing so would require longer monitoring than 24 hours and additional validation studies. Additionally, longitudinal monitoring could necessitate a specific, rather than sensitive assay, to prevent alarm fatigue driven by false positives. Prospective studies with larger sample sizes and a testable intervention will need to be performed in order to determine clinical utility.

A third limitation is our lack of locomotor activity data, which if present may have enhanced the accuracy of our classifier. Previously we have shown the addition of locomotor activity to HRV metrics improves accuracy of classification of schizophrenia (Osipov et al. 2015). This could also be the case for PTSD; locomotor activity may improve signal quality assessment or directly indicate disturbed sleep, sedentary behavior, or avoidance of traumatic stimuli.

A fourth limitation is model output being probability of a PTSD diagnosis, which is a coarse proxy for illness severity. Our method would estimate a low probability of illness for a subject who is diagnosed with PTSD yet has atypically low levels of ANS dysfunction. Other aspects of PTSD symptomatology described in the DSM-V – such as negative alterations in mood or problems concentrating – have yet to be evaluated in the context of HRV

measures. Estimating particular manifestations of PTSD severity may be more clinically useful than estimating PTSD status. However, doing so would require larger studies with multimodal data including high-resolution ECG recordings, locomotor activity, and clinical questionnaires.

Despite several limitations, this approach of classifying mental illness from physiological data has applications beyond PTSD. Changes in ANS function and psychological stress occur in other psychiatric illnesses such as bipolar disorder and depression, and are detectable using noninvasive physiological sensors (Burns et al. 2011; Sano et al. 2013; Tsanas et al. 2016; Palmius et al. 20177). Previously we used HRV measures and locomotor activity to accurately separate subjects with schizophrenia from healthy controls (Osipov et al. 2015). Our novel approach of extracting features from quiescent segments of RR intervals could also be applied to locomotor activity, which correlates with illness status and HR. Techniques that improve the signal-to-noise ratio and enable fusing of complementary data sources could aid the classification of other mental illnesses. Other possible applications of this approach are to monitor adherence to medication, or to assess efficacy of an intervention. Interpreting model output as illness severity rather than a probability of class membership could alert a caregiver of deterioration or a sustained problem in a patient.

The utility of computational approaches to interpret multiple statistical and dynamic features of physiological signals has become increasingly apparent in all fields of biomedicine. Complex, information-rich settings such as critical care or sleep medicine are especially fertile sources of data with which to build tools and address clinical questions (Monasterio et al. 2012; Behar et al. 2013).

3.6 Conclusion

We classified PTSD in 48 male veterans using L1L2 regularized logistic regression trained on HR and HRV features. Classifiers trained on the most predictive four features from 24 hours or random ten-minute control segments of RR intervals achieved test AUCs of 0.67

and 0.70, respectively. Test AUC was increased to 0.86 by segmenting RR intervals into quiescent ten-minute segments to filter out activity- or noise-related effects. This approach demonstrates the feasibility of a simple HR-based windowing approach for identifying segments of data with information correlating to the output classes predicted by the model. To our knowledge this is the first report of classification of PTSD status using non-invasive physiological features. This approach may provide a long-term ambulatory index of PTSD severity, have applications in the study and management of other mental illnesses, and be useful for other clinical disciplines where cardiovascular disease and stress are significant factors. We emphasize a passive monitoring approach would not be used for diagnosing patients in lieu of a trained clinician, and would rather be used to assess patient status in near real-time. Large randomized controlled trials comparing passive monitoring of PTSD to standard of care are required to determine the viability of this approach for clinical decision support. Lastly, the output of any monitoring system must be connected to clinically meaningful interventions.

CHAPTER 4
COMBINING HEART RATE AND LOCOMOTOR ACTIVITY DATA TO CLASSIFY SCHIZOPHRENIA

4.1 Overview

Objective. Schizophrenia has been associated with changes in HR, HRV, and locomotor activity. A previous study used HRV and locomotor activity features to accurately dichotomize patients from controls, and achieved nearly perfect classification with an AUC of 0.99 when using the ten days of data with the least missing data among all days of data (Osipov et al. 2015). However, the window length – number of contiguous days of time series data from which features are derived – dictates the time scale over which relevant phenomena are represented, and remains an under-explored topic in the field of passive sensing for health monitoring. A scale of hours to days contains information about circadian rhythms and sleep, a scale of days to weeks contains information about social drivers of behavior (i.e. cadence of the work week), and a scale of months may be necessary to measure physiology or behavior mediated by hormonal cycles or seasons. The selection of window length of data thus may be an important consideration for continuous patient monitoring.

Approach. In an effort to simulate real-time monitoring of patients diagnosed with schizophrenia, we used objective HR and activity data to classify contiguous days of data as belonging to a patient or a healthy control. HR and physical activity recordings were made on 12 medicated subjects with schizophrenia and 12 healthy controls. Features derived from these data included classical statistical characteristics, rest-activity metrics, transfer entropy, and multiscale fuzzy entropy. We varied the analysis window length from two to eight days, and selected features via minimal-redundancy-maximal-relevance. A support vector machine was trained to classify schizophrenia from control windows on a daily basis. Model performance was assessed via subject-wise leave-one-out-crossfold-validation.

*Main results.*An analysis window length of eight days resulted in an area under a receiver operating characteristic curve (AUC) of 0.96. Reducing the analysis window length to two days only lowered the AUC to 0.91. The type of most predictive features varied with analysis window length.

Significance. Our results suggest continuous tracking of subjects over short time may differentiate patients with schizophrenia from healthy controls (Reinertsen et al. 2017b). Further research is needed to determine whether this approach can accurately detect variations in symptom severity in short periods of time.

4.2 Motivation and study organization

Schizophrenia is a chronic psychiatric disease and global health problem with a lifetime prevalence of 4.0/1,000 (Saha et al. 2005). It is among the most disabling and economically catastrophic disorders; the overall cost of schizophrenia to the U.S. in 2002 was estimated at $62.7B due to clinical care, medication, and unemployment (Wu et al. 2005). Onset usually occurs in early adult years. Schizophrenia is characterized by delusions, hallucinations, disorganization of speech and behavior, and a higher rate of co-occurring psychiatric disorders. Depression prevalence is 25%, which is higher than the rate in the general population (Buckley et al. 2009), and lifetime risk of suicide is 5% (Hor et al. 2010; Kahn et al. 2015).

Schizophrenia is diagnosed via clinical interview, in which the psychiatrist asks the patient and family members about characteristic symptoms and assesses if social and/or occupational dysfunction has occurred for at least six months. However, schizophrenia impairs insight which can hinder the accuracy of self-reporting, challenging both initial diagnosis as well as subsequent monitoring of clinical status. Although treatment with medications can relieve psychotic symptoms, adherence to therapy is poor and patients generally do not achieve substantial improvements in social, cognitive and occupational functioning (Byerly et al. 2007). Cognitive-behavioral therapy, cognitive remediation, and educational and employment support are helpful but resource-intensive, as well as reliant on patient voli-

tion and/or social support systems. Given the burden of advanced illness on the patient, family members, and society, and the tendency for psychotic relapse and/or exacerbation, objective measures of clinical status could have tremendous benefit in the management of schizophrenia. Currently, psychiatric care in the United States is not delivered via a telehealth mechanism, and follow-up of clinically stable patients with schizophrenia occurs in an ambulatory outpatient setting every several weeks to months depending on severity. An accurate, passive, and objective measure of clinical status and/or severity could enable earlier identification of relapse, more rapid adjustment of medication dose to compensate for fluctuating symptoms, and identification of more detailed phenotypes of illness.

Differences in power spectral and/or entropy measures of both activity and heart rate (HR) have been reported in schizophrenia patients versus healthy controls (Bär et al. 2008; Wulff et al. 2012; Hauge et al. 2011). These sophisticated measures include frequency components and information theory complexity, contain more information than classical statistical characteristics such as the mean of a signal, and reflect changes in the ANS.

Previously we reported differences in ANS function, sleep patterns, and locomotor activity in subjects with schizophrenia versus healthy controls. We accurately distinguished patients already diagnosed with schizophrenia from healthy controls by training a machine learning algorithm with features derived from HR and locomotor activity selected from the highest-quality ten days of data recorded from a body worn patch (Osipov et al. 2015). Using both HR and activity features improved classification accuracy compared to using features from just one data type. However, the amount of data used was selected on the basis of maximizing classifier performance rather than practical study design considerations. Furthermore, time scales of relevant physiology and behavior differ. A feature calculated from a few days of data may contain information about systems mediated by circadian rhythms and sleep. Alternatively, a feature calculated from a week or more of data may contain information about social activity, behavior, and lower-frequency physiological dynamics. Assessing if feature selection and classifier performance varies with different lengths of data could improve

our understanding of how features relevant to mental illness depend on time scale. Such work could contribute to passive, objective, and near real-time monitoring of schizophrenia patients to detect early signs of illness relapse, medication adherence, or treatment efficacy. Here we vary the analysis window length of recorded HR and locomotor activity data, extract features from these data, train a support vector machine (SVM) to differentiate patients with schizophrenia from healthy controls, and evaluate classifier performance for differing analysis window lengths.

4.3 Methods

4.3.1 Participants and data collection

16 clinically stable outpatient subjects diagnosed with schizophrenia, and 19 healthy control volunteers without a history of mental disorders were recruited for the study. All subjects were unemployed. Although the prevalence of schizophrenia is only 4/1000 in the general population, this balanced cohort was appropriate for our study aim: to develop a method for estimating illness severity in subjects with an established diagnosis of schizophrenia. Age and gender did not significantly differ among the two groups, as assessed via a two-sided Student's t-test and Fisher's exact test, respectively. Subjects diagnosed with schizophrenia were taking anti-psychotic medications including Olanzapine, Risperidone, Aripiprazole, Perphenazine, Fluphenazine, Ziprasidone, Haloperidol, and Quetiapine.

HR and locomotor activity were monitored for 3-4 weeks using a disposable adhesive patch sensor worn on the chest and manufactured by Proteus Biomedical (Redwood City, CA). Electrocardiogram (ECG)-derived HR data were collected every 10 min by calculating mean HR over 15-sec intervals. Accelerometry-derived locomotor activity data were collected every 5 min by calculating mean acceleration over 15-sec intervals. Data were transmitted to a mobile phone via Bluetooth and further uploaded to a central server for processing.

4.3.2 Data pre-processing

Matlab R2016a (Mathworks, Natick, MA) was used to analyze HR (beats per minute, or BPM) and locomotor activity (normalized units $\in [0, 1]$) time series data, which were collected with variable recording rates and lengths. Data collected at an insufficient sampling rate – after an interval exceeding 1.5× the sampling period, which equals 15 min for HR and 7.5 min for activity – were discarded. Additionally, HR values lower than 20 BPM or higher than 160 BPM were labeled as low-quality and removed. A day was considered to have sufficient data if it contained both a) at least 50 HR data points and b) at least 50 locomotor activity data points. Subjects with fewer than 10 days of sufficient data were removed.

Four subjects with schizophrenia lacked sufficient amounts of data, and 12 subjects with schizophrenia possessed sufficient amounts of data. Prior to analysis, excess control subjects were removed at random to ensure both groups had 12 patients (24 subjects total). HR data were re-sampled to 10-min intervals, and activity data were re-sampled to 5-min intervals, both via linear interpolation.

Data pre-processing, feature extraction, classifier training, and model evaluation were performed for sliding contiguous windows of length of w days where $w \in [2, 4, 6, 8]$. Days with at least 50 HR and at least 50 activity data were considered viable. Days with fewer data were skipped. The $i^{\texttt{th}}$ analysis window started at day d_i and ended at day $d_i + w - 1$ where d_i is the $i^{\texttt{th}}$ day of data. i started at 1 and incremented to the last day of viable data for a subject. If a day was not viable, i.e. lacked sufficient data, no features were extracted, schizophrenia status was not estimated, and the algorithm incremented to the next day.

4.3.3 Statistical characteristics

HR and locomotor activity are affected in schizophrenia so we calculated classical statistical characteristics of HR and activity, including mean, median, mode, standard deviation (σ) and interquartile range (IQR) (Bär et al. 2008; Hauge et al. 2011; Rachow et al. 2011).

4.3.4 Rest-activity characteristics

Circadian rhythm disruption has been shown to play a significant role in schizophrenia (Wulff et al. 2012). We calculated rest–activity characteristics of HR and activity, including mean level during the least active five hours (L5), mean level during the most active 10 hours (M10), relative amplitude (RA, equation 1), interday stability (IS, Equation 2) and intraday variability (IV, Equation 3) (Witting et al. 1990; Van Someren et al. 1999).

4.3.5 Behavioral features

The same features were extracted as described in our previous work, including basic statistical characteristics, rest-activity characteristics, and multiscale transfer entropy (MTE) between HR and activity data. 36 total features were calculated, which are described in Osipov et al. 2015. However, instead of multiscale entropy (MSE) we calculated multiscale fuzzy entropy, a novel metric of sequence complexity described earlier in this thesis.

4.3.6 Multiscale fuzzy entropy

Lower time scales of MSE provide the best features for training a classifier to distinguish schizophrenic from healthy subjects (Osipov et al. 2015). Therefore we evaluated MFE of HR and activity data for the first four time scales (in the coarse-graining sense, not window length). Furthermore, a classifier trained on MFE of HR data from patients with heart failure outperformed a classifier trained on MSE of the same data (Liu et al. 2013). Thus, we calculated MFE of HR and activity data for the first four time scales, using parameter values $m_{\mathrm{HR}} = 3$, $m_{\mathrm{act}} = 2$, $r_{\mathrm{HR}} = 0.1$, and $r_{\mathrm{act}} = 0.15$.

4.3.7 Transfer entropy

Transfer entropy given by $\mathcal{T}_{X\to Y}$ is a measure of directional coupling between two concurrently sampled time series $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$. Formally, $\mathcal{T}_{X\to Y}$ is a reduction in uncertainty, given by the conditional entropy of y_i given its past values minus

the conditional entropy of y_i given both its past values and past values of the other variable $y_{i-w}^{(l)}$:

$$\begin{aligned}\mathcal{T}_{X\to Y} &= H(y_i|y_{i-w}^{(l)}) - H(y_i|y_{i-w}^{(l)}, x_{i-t}^{(k)}) \\ &= \sum_{y_i, y_{i-w}^{(l)}, x_{i-t}^{(k)}} p(y_i, y_{i-w}^{(l)}, x_{i-t}^{(k)}) \log \frac{p(y_i|y_{i-w}^{(l)}, x_{i-t}^{(k)})}{p(y_i|y_{i-w}^{(l)})} \\ &= \sum_{y_i, y_{i-w}^{(l)}, x_{i-t}^{(k)}} p(y_i, y_{i-w}^{(l)}, x_{i-t}^{(k)}) \\ &\quad \log \frac{p(y_i, y_{i-w}^{(l)}, x_{i-t}^{(k)}) p(y_{i-w}^{(l)})}{p(y_{i-w}^{(l)}, x_{i-t}^{(k)}) p(y_i, y_{i-w}^{(l)})}\end{aligned} \tag{4.1}$$

where i indicates a given point in time, t and w are the time lags in X and Y respectively, and k and l are the block lengths of past values in X and Y respectively. k and l were both set to 1 to improve computational speed (Lee et al. 2012).

Multiscale transfer entropy (MTE) was calculated by coarse-graining HR and activity time series τ times, estimating joint probability distribution functions via D-V partitioning, and calculating $\mathcal{T}_{\mathrm{HR}\to\mathrm{act}}(\tau)$ and $\mathcal{T}_{\mathrm{act}\to\mathrm{HR}}(\tau)$ for $\tau = 1, 2, ..., \tau_{\max}$ time scales.

4.3.8 Feature selection

In order to reduce the number of features used and minimize overfitting, a feature selection approach was necessary. Features were z-scored, discretized, and ranked via minimum Redundancy Maximum Relevance (mRMR) criteria (Peng et al. 2005). This approach simultaneously minimizes mutual information between individual features x in a feature set S:

$$\min\ R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \tag{4.2}$$

and maximizes mutual information between features x and classes (sometimes referred to as labels or targets) c:

$$\max D(S,c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{4.3}$$

where $I(x;y)$ is the mutual information between variables x and y, and $p(x)$, $p(y)$, and $p(x,y)$ are probability densities of these variables:

$$I(x;y) = \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{4.4}$$

These two constraints for D and R are combined into one expression $\phi(D,R)$ which is optimized as follows:

$$\max \phi(D,R) = D - R \tag{4.5}$$

The mRMR algorithm ordered features by values of ϕ from highest to lowest. Hereafter, "most predictive" refers to the subset of features with the highest values of ϕ.

Statistical characteristics, rest-activity characteristics, MTE, and MFE were calculated and 36 features total were ranked via mRMR. The most predictive i features (where $i \in 1, 2, ..., m$) were used to train a machine learning algorithm.

4.3.9 Classification of schizophrenia status among subjects

Subjects were classified as either having a diagnosis of schizophrenia or being healthy, using `libsvm`, an open-source SVM library (Chang et al. 2011). The two–dimensional matrix of features consisted of W windows by m features, and the one-dimensional array of labels consisted of W binary labels. Each analysis window was labeled as 1 if belonging to a schizophrenia patient, or 0 if belonging to a healthy control. A Gaussian radial basis function kernel with $\gamma = 0.0312$ was selected based on previous work (Osipov et al. 2015).

The most predictive features identified via mRMR were used to train the SVM, which output probability estimates of a subject being labeled as schizophrenic, $P(SZ)$, rather than healthy. Classifier performance was assessed via subject-wise leave-one-out crossfold

validation (LOOCV; given N patients, $N-1$ patients are used to train the classifier and the remaining patient is used as the test set), and various attributes of classifier performance were calculated, including area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The number of most predictive features (number of features maximizing the AUC) was also determined.

Pre-processing, feature extraction, and classifier assessment were performed for analysis window lengths of two, four, six, and eight days. The SVM was trained using windows, not on individual days or subjects.

4.4 Results

The number of features resulting in the maximum AUC, and the maximum AUC value itself, both differed with analysis window length (Figure 4.1). With a analysis window length of two days, the model achieved a maximum AUC of 0.91 using three most predictive features. With an analysis window length of eight days, the model achieved a maximum AUC of 0.96 using 11 most predictive features. Table 4.1 lists other classifier performance metrics for varying window lengths.

Box plots of the most predictive features (i.e. the combination of features which maximized the training AUC) for two-day and eight-day analysis windows are shown in Figure 4.2 and Figure 4.3 respectively. For two-day analysis windows, the three most predictive features in order of more to less predictive are 1) the standard deviation of activity (σ_{act}), 2) the multiscale fuzzy entropy of heart rate at the first time scale ($\mathrm{MFE}_{\mathrm{HR},1}$), and 3) the mode of activity (Mo_{act}).

For eight-day analysis windows the 11 most predictive features in order of more to less predictive are 1) $\mathrm{MFE}_{\mathrm{HR},1}$, 2) IQR_{act}, 3) Mo_{act}, 4) the multiscale transfer entropy from activity to HR at the first time scale (coarse-graining) ($\mathrm{MTE}_{\mathrm{act}\rightarrow\mathrm{HR},1}$), 5) σ_{act}, 6) $\mathrm{MFE}_{\mathrm{act},4}$, 7) $\mathrm{MFE}_{\mathrm{HR},4}$, 8) the intraday variability of activity (IV_{act}), 9) $\mathrm{MTE}_{\mathrm{act}\rightarrow\mathrm{HR},2}$, 10) $\mathrm{MFE}_{\mathrm{act},2}$,

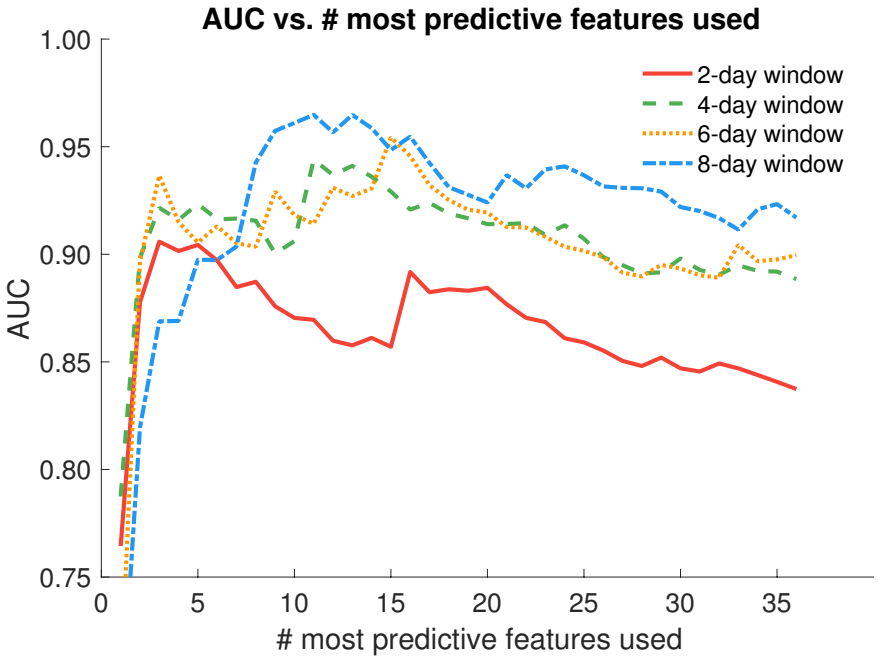


Figure 4.1: AUC versus number of most predictive features, selected out of 36 total features via mRMR, used to train the SVM. The blue line represents two-day analysis windows and the red line represents eight-day analysis windows. The maximum AUC for two-day analysis windows is 0.91 using the three most predictive features, and the maximum AUC for eight-day analysis windows is 0.96 using the 11 most predictive features.

and 11) the relative amplitude of activity (RA_{act}).

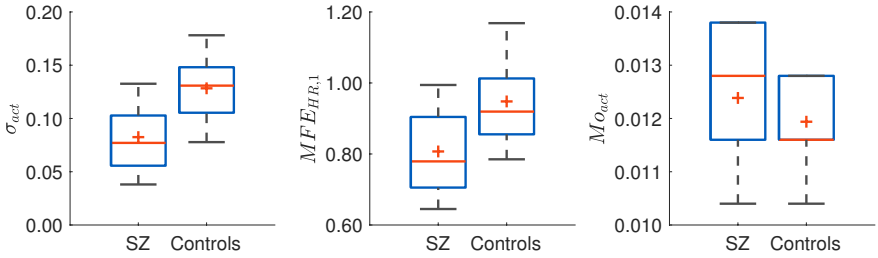


Figure 4.2: Box plots of most predictive features selected via mRMR using two-day analysis windows. The SZ label on the x-axis indicates features from schizophrenia patients. These three features in combination maximized the training AUC. The red + indicates the mean, the middle horizontal red line indicates the median, the blue box denotes the interquartile range (IQR) flanked by the 25th and 75th percentiles, and the vertical lines outside of the box indicate the 9th and 91st percentiles. The median value of every feature significantly differed by schizophrenia status, with $P < 0.05$ calculated via two-sided Wilcoxon rank-sum test.

Probability density estimates of classifier output (estimated probability of a window of data belonging to a subject with schizophrenia) for schizophrenia patients distinctly differed from those of control subjects. This difference was large for both two-day (Figure 4.4a) and eight-day analysis windows (Figure 4.4b).

ROC curves showed a positive correlation between analysis window length and AUC (Figure 4.5). AUC values ranged from 0.91 for two-day analysis windows to 0.96 for eight-day windows. Increasing window length increased most metrics of classifier performance (Table 4.1) with only a few exceptions. Increasing window length from two to four days reduced sensitivity from 0.89 to 0.85 and NPV from 0.79 to 0.76. Increasing analysis window length from four to six days reduced specificity from 0.98 to 0.94, and PPV from 0.98 to 0.97. Increasing window length from six to eight days reduced NPV from 0.77 to 0.76.

To assess the relative contribution of each type of feature (i.e. heart rate, locomotor activity, or both combined) we compared classifier AUC versus feature type for analysis window lengths of two or eight days. With an analysis window length of two days, features from HR resulted in a classifier AUC of 0.80, features from locomotor activity resulted in a classifier AUC of 0.85, and features from both HR and locomotor activity resulted in

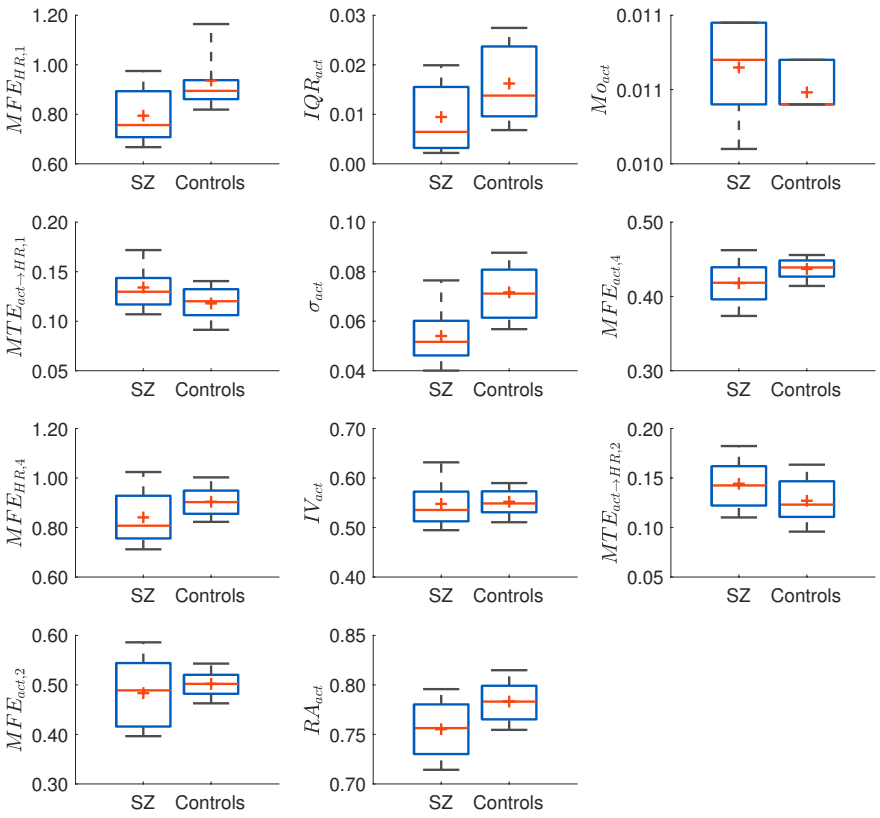
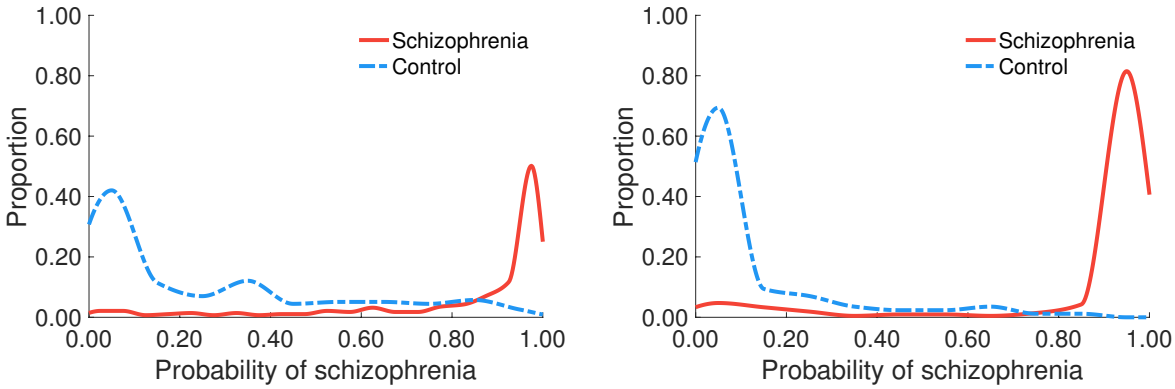


Figure 4.3: Box plots of most predictive features selected via mRMR using eight-day analysis windows. The SZ label on the x-axis indicates features from schizophrenia patients. These 11 features in combination maximized the training AUC. The red + indicates the mean, the middle horizontal red line indicates the median, the blue box denotes the interquartile range (IQR) flanked by the 25th and 75th percentiles, and the vertical lines outside of the box indicate the 9th and 91st percentiles. The median value of every feature significantly differed by schizophrenia status, with $P < 0.05$ calculated via two-sided Wilcoxon rank-sum test.



(a) Two-day analysis windows

(b) Eight-day analysis windows

Figure 4.4: Probability density estimates of classifier output – estimated probability of a window of data belonging to a subject with schizophrenia – using (a) two-day and (b) eight-day analysis windows. Leave-one-out cross-validation was performed. Classifier output is on the x-axis, and proportion is on the y-axis; all y-values for a class sum to unity.

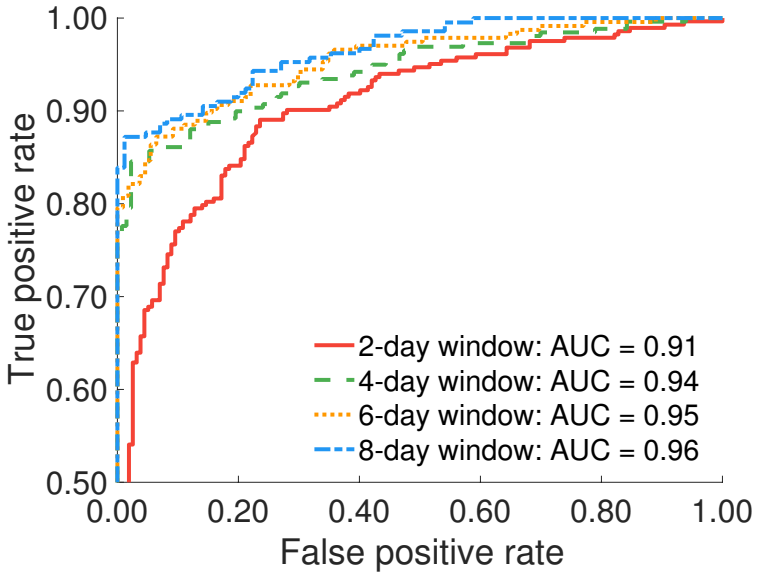


Figure 4.5: Receiver operating characteristic (ROC) curves vary with analysis window length. Leave-one-out cross-validation was performed. Blue, red, yellow and purple denote two, four, six, and eight-day windows respectively. The y-axis is the true positive rate, or sensitivity. The x-axis is the false positive rate, or 1 − specificity.

a classifier AUC of 0.91 (Table 4.2). These results demonstrate an improvement classifier performance when using features from both HR and locomotor activity data, compared to using features from either category alone.

HR and activity data, estimated $P(SZ)$, optimal classifier thresholds, and data quantity were visualized against 24-hour intervals into the study for a representative subject with schizophrenia (Figure 4.6a) and a representative healthy control subject (Figure 4.6b).

Table 4.1: Classifier performance metrics versus window length, using both HR and activity features. Leave-one-out cross-validation was performed. Reported metric is for held-out test set data. PPV indicates Positive Predictive Value and NPV indicates Negative Predictive Value.

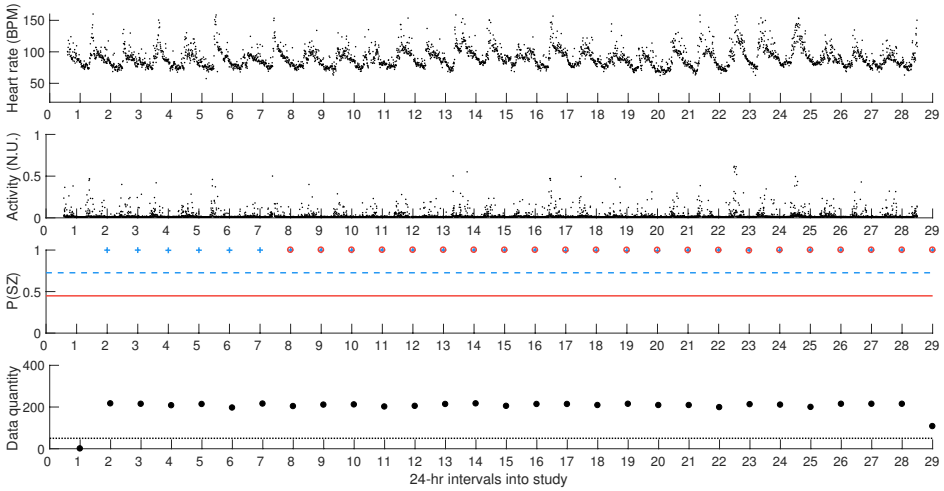
	Window length (days)			
Metric	2	4	6	8
AUC	0.91	0.94	0.95	0.96
Accuracy	0.85	0.89	0.89	0.91
Sensitivity	0.89	0.85	0.87	0.87
Specificity	0.76	0.98	0.94	0.99
PPV	0.87	0.98	0.97	0.99
NPV	0.79	0.76	0.77	0.76

Table 4.2: Area under the ROC curve (AUC) vs. window length and feature type used to train support vector machine. Leave-one-out cross-validation was performed. AUCs reported for held-out test set data, calculated via leave-one-out-cross-validation.

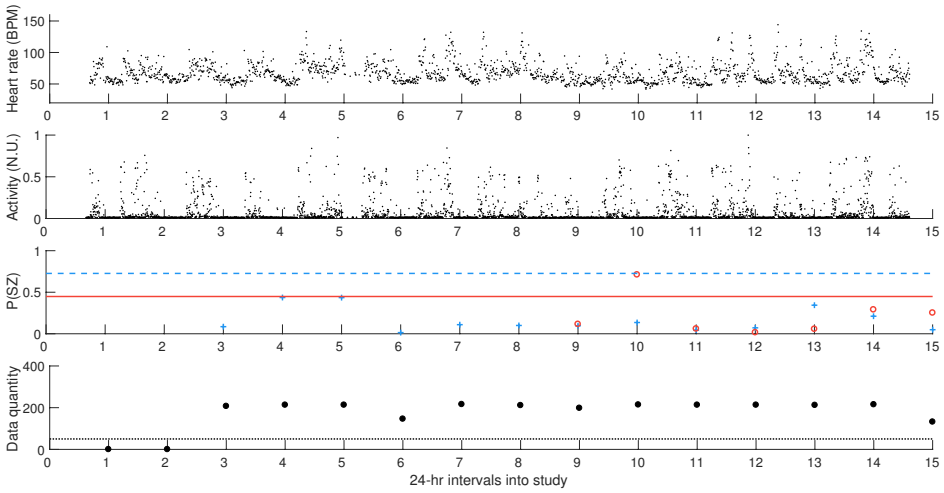
	Window length (days)	
Feature type(s)	2	8
Heart rate (HR)	0.84	0.90
Activity	0.86	0.89
HR and activity	0.91	0.96

4.5 Discussion

We built on previous work using features derived from HR and activity to classify medicated schizophrenia patients from healthy controls. Here we evaluated the relationship between analysis window length and classifier accuracy.

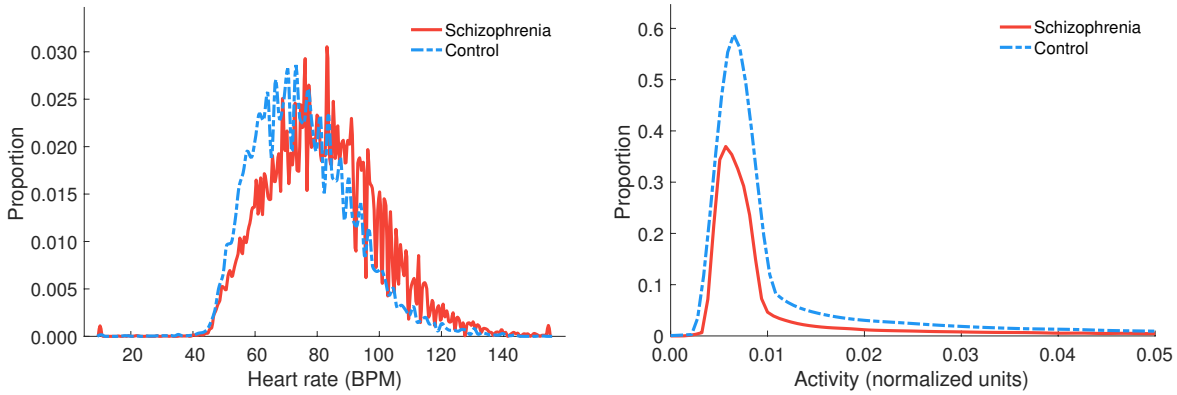


(a) Subject with schizophrenia



(b) Healthy control subject

Figure 4.6: HR data (top row), activity data (second row), classifier output (probability of schizophrenia, or $P(SZ)$; third row), and data quantity versus time (bottom row) for a (a) schizophrenia patient and a (b) healthy control subject. Heart rate is in beats per minute (BPM), activity is in normalized units (N.U.), $P(SZ)$ is a probability, and data quantity is in raw counts. $P(SZ)$ for a window length of two days is shown by the red +'s. The classifier threshold for a two-day window is $P(SZ) = 0.45$, is shown by the red solid line. $P(SZ)$ for a window length of eight days is shown by the blue circles. The classifier threshold for a window length of eight days, $P(SZ) = 0.73$, is shown by the blue dashed line. On the data quantity plot, the minimum data quantity (at least 50 HR and at least 50 activity data) required to make a estimate of $P(SZ)$ on a given day is shown by the black dotted line.



(a) HR data from all subjects

(b) Activity data from all subjects

Figure 4.7: Distribution of raw a) HR data and b) activity from all subjects, separated by schizophrenia status. Distributions of HR data were significantly different (P¡0.001; two-sample Kolmogorov-Smirnov test), but medians were not significantly different (P=1.00; two-sided Wilcoxon rank sum test). Results were identical when these tests were performed on activity data.

Using two-day analysis windows, a maximum AUC of 0.91 was achieved using the three most predictive features. Using eight-day analysis windows, a maximum AUC of 0.96 was achieved using the 11 most predictive features (Figure 4.1). Both AUCs sharply increased as initial features were added. For the two-day model, adding more than three features steadily lowered the AUC, aside from a slight rise in classifier performance when moving from 16 to 17 features. On the other hand, for 4-, 6-, and 8-day models the AUC stayed close to 0.90 for a wide range of 10 to 25 features. This may be due to the fact that, for the two-day model, the number of data points is low enough that higher-dimensional models begin to fit more noise.

The ranking of features via mRMR criteria depended on analysis window length (4.2, 4.3). The most predictive feature from two-day analysis windows was σ_{act} with a median $\pm$ IQR of 0.08 ± 0.05 units for the schizophrenic group compared to 0.13 ± 0.04 units for the control group. This difference was statistically significant ($P < 0.05$, Wilcoxon rank-sum test), contrasting with results from Hauge et al. 2011, which also reported significantly lower levels of mean activity in schizophrenic patients. Our algorithm did not identify mean activity as a predictive feature. However, that work reported the evaluation of long-term

patients from an open ward, whereas our study evaluated outpatients in relative symptomatic remission. Illness severity and behavioral markers may have differed.

The 2nd most predictive feature for two-day analysis windows was $\mathrm{MFE}_{\mathrm{HR},1}$, with a median $\pm$ IQR of 0.95 ± 0.50 in schizophrenic patients and 1.30 ± 0.40 in controls. The 3rd most predictive feature for two-day analysis windows was Mo_{act}, with a median $\pm$ IQR of 0.013 ± 0.002 in schizophrenic patients and 0.012 ± 0.001 in controls. All three most predictive features for two-day analysis windows differed significantly by schizophrenia status ($P < 0.05$, Wilcoxon rank-sum test).

The most predictive feature for eight-day windows was $\mathrm{MFE}_{\mathrm{HR},1}$, with a median $\pm$ IQR of 0.89 ± 0.46 for schizophrenic patients and 1.24 ± 0.19 for controls (4.3). This difference was consistent and statistically significant for both two- and eight-day analysis windows ($P < 0.05$, Wilcoxon rank-sum test). Other predictive features for eight-day analysis windows represent the spread of the distribution of activity, such as IQR_{act} and σ_{act}. We note skewness, kurtosis, mean, or median of activity were not predictive features. Furthermore, aside from $\mathrm{MFE}_{\mathrm{HR},1}$ and $\mathrm{MFE}_{\mathrm{HR},4}$, no other HR statistics were predictive.

$\mathrm{MTE}_{\mathrm{act}\to\mathrm{HR},i}$ denotes the transfer entropy, or amount of uncertainty reduced in future values of HR by knowing the past values of activity given past values of HR, at a time scale i. For two-day analysis windows, $\mathrm{MTE}_{\mathrm{act}\to\mathrm{HR},i}$ was not a predictive feature for any time scale. However, when using eight-day analysis windows, $\mathrm{MTE}_{\mathrm{act}\to\mathrm{HR},i}$ was a predictive feature for $i = 1$ and $i = 2$. When using two-day analysis windows, $\mathrm{MTE}_{\mathrm{act}\to\mathrm{HR},i}$ was still predictive; however, it was not as predictive as the most predictive three features described above. These results suggest activity-driven changes in HR become more predictive at longer time scales.

We used the mRMR framework to select features on the basis of maximizing predictivity (relevance), and minimizing collinearity (redundancy) (Peng et al. 2005). Although our most-predictive features were all separable in a univariate sense, features should be selected by predictivity rather than significance for classification tasks (Lo et al. 2015). Features

may be inseparable in a univariate sense but strongly separable in higher dimensions. On the other hand, a feature may significantly differ between classes in a univariate or even multivariate sense, but may still not contribute predictive accuracy to a model.

With two-day analysis windows, the SVM classifier produced grossly distinct distributions of output for schizophrenia patients versus healthy controls, demonstrating excellent classifier performance (Figure 4.4a). However, a few analysis windows were mislabeled. Using eight-day analysis windows resulted in similar excellent classifier performance evidenced by large separability of classifier output between windows of data from schizophrenia patients and healthy controls (Figure 4.4b).

AUC increased with analysis window length, from 0.91 for two days, up to 0.96 for eight days (Figure 4.5). In our previous work, an SVM was trained using the 10 days missing the least data from each subject, and achieved an AUC of 0.99 (Osipov et al. 2015). In this work, classifier performance was slightly lower because we did not pick "best" days among several weeks of data. Rather, we slid an analysis window through all days from all patients and trained a classifier on individual windows. This approach better represents a realistic scenario in which only a few days of patient data are obtained.

We also assessed how analysis window length affects the sensitivity, specificity, positive predictive value, negative predict value, and accuracy of the classifier for different window lengths (Table 4.1). Specificity exceeded 0.90 for four-, six-, and eight-day windows. The AUC uniformly increased with window length. Compared to two-day windows, four-day windows resulted in a higher AUC, accuracy, and specificity, but sensitivity dropped from 0.89 to 0.85. A decrease in sensitivity is undesirable when the cost of missing a true positive is high (*e.g.* suicide). A monitoring system to track illness severity of patients already diagnosed with schizophrenia might thus prioritize sensitivity, although doing so increases false positives.

Analysis window length determined which features are most predictive, mediates classifier performance, perhaps due to differing time scale of relevant physiology and behavior. Prior

work on the association between schizophrenia and disturbances in the ANS as measured by heart rate variability metrics has shown a similar dependence on physiologically and behaviorally relevant time scales (Bär et al. 2008; Rachow et al. 2011). A feature calculated from two days of data may contain information about systems mediated by circadian rhythms and sleep. Alternatively, a feature calculated from eight days of data may contain information about social activity, behavior, and lower-frequency physiological dynamics. On the scale of days to weeks, social drivers of behavior (i.e. cadence of the work week) that are not present on the scale of hours may become more apparent in these data. Schizophrenia patients typically have disrupted social routines (Kahn et al. 2015). Recordings on the order of months may be necessary to measure physiology or behavior mediated by hormonal cycles or seasons. The selection of analysis window length thus may be an important consideration in the design of studies for monitoring patients with mental illness.

We assessed the relative contribution of HR and locomotor activity features to classifier accuracy. Previously we have shown a combination of HR and activity features outperforms either HR or activity features alone (Osipov et al. 2015). Here we show the same trend for both two- and eight-day analysis windows (Table 4.2). For two-day analysis windows, using activity features (AUC of 0.86) resulted in better classifier performance than HR features (AUC of 0.84). For eight-day analysis windows, using HR features (AUC of 0.90) resulted in slightly better classifier performance than activity features (AUC of 0.89). Using both HR and activity features together improved the AUC, for both two- and eight-day analysis windows. Locomotor activity and HR are correlated – the former introduces artifact and random error to HR, and HR tends to rise during locomotor activity – yet contribute complementary and non-redundant information about subject behavior that improves the predictive accuracy of a classifier.

Individual data for patients, estimated probabilities of schizophrenia $P(SZ)$, and data quantity for each 24-hr interval were visualized for a representative schizophrenia patient (4.6a) and healthy control subject (4.6b). Time series data were re-sampled during signal

processing, so overall data quantity per 24-hr interval was almost equal between schizophrenia and control subjects. Upon visual inspection, HR appears more periodic for the schizophrenia patient compared to the control. This observation is consistent with $\mathrm{MFE}_{\mathrm{HR},1}$ being significantly lower in schizophrenia patients than in controls for both two- (4.2) and eight-day analysis windows (4.3), and with previous reports of lower HR complexity in these patients (Rachow et al. 2011).

For some schizophrenia patients, estimated $P(SZ)$ occasionally fell below the classification threshold (Figure 4.6a). Likewise, for some control subjects, estimated $P(SZ)$ rose above the threshold (Figure 4.6b). Due to the lack of more richly labeled data, it is unclear if this misclassification indicates control subjects have schizophrenic-like days and schizophrenia patients have normal-like days. Regardless, averaging strategies could theoretically reduce false positives. The optimal trade-off between sensitivity or specificity is determined by the use case and cost of false positives or false negatives; here we simply maximized the AUC.

We note several limitations of our study. Our sample size was relatively small and limited to patients from one geographic region and institution. Also, several factors such as mental and behavioral state, social status, and medication usage affect HR and activity.

Although we controlled for employment status as a proxy for social routine – which relates to activity and restfulness to some extent – we did not explore other potential confounders that could affect HRV or locomotor activity. Literature suggests potential confounders such as weight, BMI, diet, smoking status, renal function, etc. have a moderate effect on HRV. However, the dominant factor by an order of magnitude is the mental response (Bernardi et al. 2000). The second-largest dominant factor is physical movement (Knoepfli-Lenzin et al. 2010). In this work we also analyzed locomotor activity, which we have previously shown contains predictive information for classifying subjects by schizophrenia status (Osipov et al. 2015).

Stress relates to mental state with physiological and behavioral manifestations, and could thus influence our features. Skin conductance, a biomarker for stress mediated by the sym-

pathetic nervous system, has been shown to differ in schizophrenia patients (Bär et al. 2008). Additionally, cortisol secretion and stress sensitivity may be associated with schizophrenia, or subsequent development of schizophrenia following the prodromal phase of the illness (Walker et al. 2013; Holtzman et al. 2013). Stress affects activity as well as other aspects of mobile device usage; Sano et al. 2013 reported using using screen on, mobility, call or activity level information to distinguish stressed from non-stressed individuals with an accuracy of 75%.

Aside from matching employment status, we did not directly control for stress in our study. Doing so with conscious patients is challenging especially in an ambulatory setting. Additionally, stress invariably accompanies social risk factors such as physical abuse, sexual abuse, maltreatment and bullying, which are associated with increased risk of later schizophrenia (Stilo et al. 2010).

Evaluating data solely from periods of lower activity or stress based on physiology could reduce confounding from these random variables. Recently we attempted to reduce noise in data and improved classifier performance by selecting quiescent periods of data with lowest median HR, which is a proxy criterion for restfulness (Reinertsen et al. 2017a). More sophisticated change point detection approaches could potentially sort data into parametrically similar segment, and even further increase the signal-to-noise ratio of features derived from data within each segment (Adams et al. 2007).

Antipsychotic medications have been reported to exacerbate ANS dysfunction in schizophrenia patients, potentially putting patients at greater risk of cardiac mortality (Rechlin et al. 1994; Birkhofer et al. 2013; Huang et al. 2013). On the other hand, Mondelli et al. demonstrated that antipsychotic medication can reduce cortisol secretion and normalize HPA-axis hyperactivity in patients suffering from psychosis (Mondelli et al. 2010). Bär et al. 2008 did not find significant changes in ANS function after antipsychotics were administered to patients, while Henry et al. 2010 et al. found that risperidone, valproate, or mood stabilizers did not significantly affect HRV in bipolar or schizophrenia patients. We lacked more de-

tailed information about the type, dose, or adherence of antipsychotic medications taken by patients in our study. We thus cannot claim our results generalize to non-medicated subjects with schizophrenia. However, a classifier affected by a patient's medication could potentially be used to monitor adherence and treatment efficacy.

Illness severity in schizophrenia fluctuates from day to day (Kahn et al. 2015) and strongly correlates with measures of ANS dysfunction such as HR variability (Henry et al. 2010; Montaquila et al. 2015; Bär et al. 2005; Bär et al. 2008). Our classifier output varied from day to day (Figure 4.4a), was based on measures of ANS dysfunction and behavior, and may have reflected fluctuations in illness severity. However, we lacked detailed information about daily changes in symptoms, e.g. Brief Psychiatric Rating Scale survey data, which would be necessary for training a classifier to estimate illness severity accurately.

Lastly, some hyper-parameters of our model, such as entropy template length and minimum amount of data per day, were selected from prior work instead of optimized. Classifier performance could likely be improved using techniques such as Bayesian optimization (Ghassemi et al. 2014; Shahriari et al. 2016).

4.6 Conclusion

We evaluated the relationship between analysis window length of data recordings and classifier accuracy in a small cohort of schizophrenia patients and healthy controls. A support vector machine was trained on HR and locomotor activity data obtained via body-worn patches and windowed over varying lengths ranging from two to eight days. A novel metric – multiscale fuzzy entropy – contributed to the predictive accuracy of our model. Our approach accurately classified schizophrenia status in a small cohort of subjects with an AUC of 0.91 for an analysis window length as short as two days, and an AUC of 0.96 for an analysis window length of eight days. Features selected as most predictive also varied with analysis window length. This work serves as technical proof of feasibility of using HR and activity features to differentiate patients with schizophrenia from healthy controls. Further

research is needed to determine whether this approach can accurately detect variations in symptom severity in short periods of time.

CHAPTER 5
INTERACTIONS BETWEEN HEART RATE AND LOCOMOTOR ACTIVITY

5.1 Overview

Objective. Changes in heart rate (HR) and locomotor activity reflect changes in autonomic physiology, behavior, and mood. These systems may involve interrelated neural circuits that are altered in psychiatric illness, yet their interactions are poorly understood. We hypothesized interactions between HR and locomotor activity could be used to discriminate patients with schizophrenia from controls, and would be less able to discriminate non-psychiatric patients from controls.

Approach. At least ten days of contiguous HR and locomotor activity were recorded via wearable patches in 16 patients with schizophrenia and 19 healthy controls. Measures of signal complexity and interactions were calculated over multiple time scales, including sample entropy, mutual information, and transfer entropy. A support vector machine was trained on these features to discriminate patients from controls. Additionally, time series were converted into a network with nodes comprised of HR and locomotor activity states, and edges representing state transitions. Graph properties were used as features. Leave-one-out cross validation was performed. To compare against non-psychiatric illness, the same approach was repeated in 41 patients with atrial fibrillation (AFib) and 53 controls.

Main results. Network features enabled perfect discrimination of schizophrenia patients from controls with an areas under the receiver operating characteristic curve (AUC) of 1.00 for training and test data. Other bivariate measures of interaction achieved lower AUCs (train 0.98, test 0.96), and univariate measures of complexity achieved the lowest performance. Conversely, interaction features did not improve discrimination of AFib patients from controls beyond univariate approaches.

Significance. Multiscale network dynamics quantified interactions between HR and locomotor activity. These features enabled perfect discrimination of subjects with schizophrenia from controls, but were less performant in a non-psychiatric illness. This is the first quantitative evaluation of interactions between physiology and behavior in patients with psychiatric illness.

5.2 Motivation and study organization

Schizophrenia is a severely disabling and chronic mental illness which affects over 21 million people worldwide (Saha et al. 2005). Currently schizophrenia is diagnosed and managed by mental health professionals, whose availability is often scarce, particularly in low- and middle-income countries (Saxena et al. 2007). Additionally, in the context of stable schizophrenia treated in the outpatient setting, months to years can pass between clinical visits despite changes in patient status over shorter time scales (NICE guideline (CG178) 2014).

To assess clinical status more frequently and without direct observation, noninvasive technologies such as smartphones and wearable devices that measure locomotor activity via accelerometry, heart rate (HR), and other signals have been investigated. High resolution accelerometry has also been used to measure changes in social routine and circadian rhythms in mental illnesses such as depression, bipolar disorder, and schizophrenia (Van Someren et al. 1999; Reinertsen et al. 2017b; Millar et al. 2004; Berle et al. 2010). Pulse sensors and electrocardiography (ECG) are used to assess HR and heart rate variability (HRV) measures, which indicate dysfunction in the autonomic nervous system (ANS). These tools could alert providers to a change in a patient's condition, monitor the effectiveness of interventions, and identify mediators of illness severity (Steinhubl et al. 2015).

While HR and locomotor activity have been assessed in a univariate sense, measures of interaction between these signals have not yet been explored, and may incrementally improve disease classification. In theory, information is transferred between cardiovascular physiology and behavior over several time scales. In normal individuals, circadian rhythms

mediate the increase in blood pressure and heart rate during the early morning prior to an increase in consciousness, which in turn leads to waking and locomotor activity in the morning. Conversely, the rising of an individual from a chair leads to a rise in blood pressure, heart rate, and sympathetic tone. These responses, interactions, and transitions between physiological and behavioral states vary over time scale, and are partially mediated by the baroreflex and central command, a feed-forward neural mechanism that contributes to motor and cardiovascular function during arousal and exercise (Hall 2010). Since patients with schizophrenia have ANS dysfunction, their physiological and physical responses to changes in HR and locomotor activity may be abnormal (Chang et al. 2009; Montaquila et al. 2015; Alvares et al. 2016).

Interactions between time series can be quantified via mutual information and transfer entropy. However, such measures are often calculated using an entire time series and fail to capture transitions between physiological and behavioral states that could provide clinically useful information. To better assess these dynamics, a time series can be evaluated over multiple time scales, and can also be represented as a network. Attributes of this network may provide a more nuanced measure of system complexity. Previously, networks have been constructed using beat-to-beat intervals from ECGs (known as RR intervals) of patients with congestive heart failure, and visually contrasted with networks from healthy controls (Campanharo et al. 2011). Recently we demonstrated the utility of network representations for the early prediction of sepsis, indicating an association between systemic inflammation and a loss of information flow between HR and blood pressure (Shashikumar et al. 2017a). To our knowledge, interactions between different signals such as HR and locomotor activity have never been quantified in patients with mental illness. Measures of these interactions could have clinical utility if found to correlate with disease status or symptom severity.

We hypothesized that measures of interactions between HR and locomotor activity over multiple time scales can be used to distinguish patients with schizophrenia from healthy controls. We also evaluated the additional predictive power of interaction measures for dif-

ferentiating between patients with schizophrenia and controls. To better understand the generalizability and limitations of this approach, we also tested if such interactions could differentiate atrial fibrillation (AFib) from sinus rhythm when applied to 10-minute wrist-band pulse and locomotor activity recordings from quietly seated subjects. This contrasted from the schizophrenia analysis because AFib is cardiac-specific, and the recordings occurred in a highly controlled setting, which limits the ability to assess behavior. We thus hypothesized that measures of interaction between HR and locomotor activity are less useful in discriminating disease status in AFib than in schizophrenia.

5.3 Methods

The overall approach is illustrated in Figure 5.1. The data and preprocessing are described in the following four subsections, followed by the extracted features in the subsequent seven subsections, and finally the classifier in the last subsection.

5.3.1 Schizophrenia study: participants and data collection

16 clinically stable and medicated outpatient subjects diagnosed with schizophrenia, and 19 healthy control volunteers without a history of mental illness were recruited for the study (previously described by Osipov *et al.* (Osipov et al. 2015). Subjects were unemployed. Age and gender did not significantly differ among the two groups, as assessed via a two-sided Student's t-test and Fisher's exact test, respectively. HR and locomotor activity were monitored for 3-4 weeks using a disposable adhesive patch sensor worn on the chest and manufactured by Proteus Biomedical (Redwood City, CA). ECG-derived HR data were collected every 10 min by calculating mean HR over 15 sec intervals. Locomotor activity data were collected every 5 min by calculating mean acceleration over 15 sec intervals. Data were transmitted to a mobile phone via Bluetooth and uploaded to a server for processing.

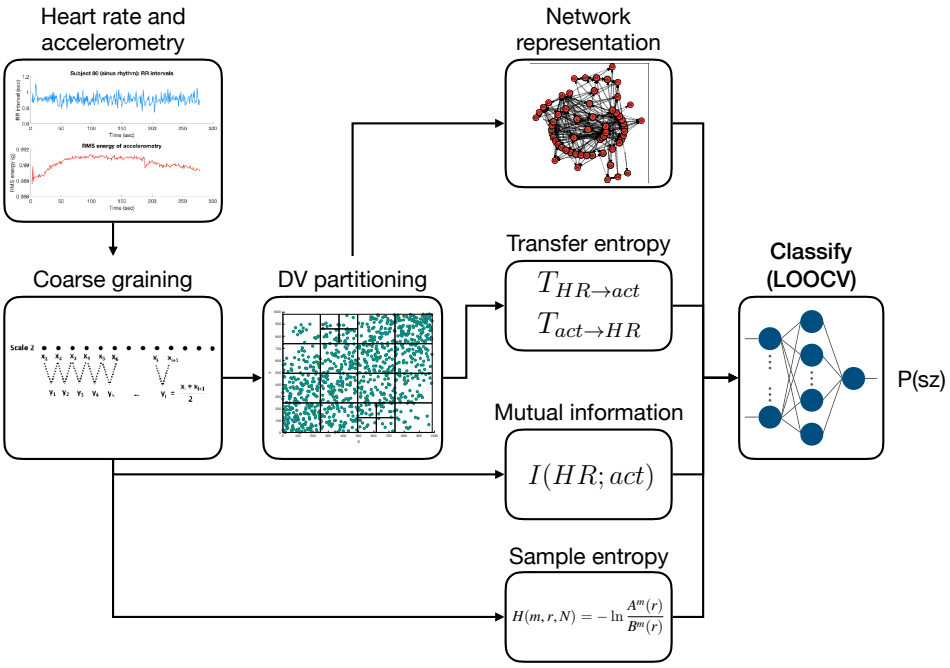


Figure 5.1: Schematic of data processing and classification algorithm. DV partitions are computed from time-lagged and coarse-grained HR and locomotor activity, which are transformed to a network representation. Topological attributes of the networks are used as input features to a machine learning classifier. DV partitions are also used to compute transfer entropy for between HR and locomotor activity (and vice-versa) for varying lags and time scales. Finally, mutual information and sample entropy are calculated for varying time scales. These features are used to train a classifier to estimate the probability of a subject belonging to the unhealthy class, $P(sz)$.

5.3.2 Schizophrenia study: data pre-processing

Data points in the time series of HR (measured in beats per minute, or BPM) and locomotor activity (measured in normalized units $\in [0, 1]$) with an interval exceeding $1.5\times$ the average sampling period, which equaled 15 min for HR and 7.5 min for activity data – were discarded. Additionally, HR values lower than 20 BPM or higher than 160 BPM were labeled as low-quality and removed. HR and activity data were re-sampled to 5 min intervals via linear interpolation. The root mean square (RMS) energy of acceleration for the i^{th} interval was calculated (equation 5.1).

Matlab R2017a (Mathworks, Natick, MA) was used for data pre-processing, feature extraction, machine learning classification, and data visualization.

5.3.3 AFib study: participants and data collection

97 subjects recruited for the study were adult patients (18-89 years old) who were hospitalized and undergoing telemetry monitoring at Emory University Hospital, Emory University Hospital Midtown, and Grady Memorial Hospital (previously described by Shashikumar *et al.* 2017 (Shashikumar et al. 2017b). The study was approved by the institutional review board of each hospital. Patients were recruited at random with an over-sampling of patients with AFib; rhythms were reviewed by an ECG technician, physician study coordinator, and cardiologist. 44 subjects had AFib and 53 had other rhythms. Three subjects with AFib had insufficient data to generate subsequent features and were excluded from analysis. Eight channel multi-wavelength photoplethysmography (PPG) and tri-axial accelerometry (x, y, z) were recorded simultaneously at a sampling frequency f_s of 125 Hz for 5 min using a research version of the wrist-worn Simband device (Samsung, Seoul, South Korea).

5.3.4 AFib study: data pre-processing

PPG data from a green light wavelength (520-535 nm) were selected because the commercially available version of the Simband contains green light sensors. Data were de-trended,

outliers greater than the 95^{th} or less than the 5^{th} percentile were removed, and a 41^{st} order bandpass filter was used with passband 0.0008 - 0.04 Hz. RR intervals were estimated from minima of peaks in the cleaned PPG data, non-physiological RR intervals greater than 2 sec or less than 0.375 sec were removed, and RR intervals occurring less than the sampling period ($1/fs$ sec) after the prior data point were removed.

5.3.5 RMS energy of acceleration

The RMS energy of acceleration during the i_{th} segment of RR intervals is given by

$$\text{RMS energy} = \sqrt{\frac{\vec{x^2} + \vec{y^2} + \vec{z^2}}{N}} \tag{5.1}$$

where $\vec{x}$, $\vec{y}$, and $\vec{z}$ are x, y, and z-axis accelerometry values in the i^{th} segment, and N is the number of accelerometry data within this segment.

5.3.6 Statistical moments

The mean, median, mode, variance, skewness, and kurtosis of HR and activity were calculated for both schizophrenia and AFib groups.

5.3.7 Varying time scales via coarse-graining

Interactions between physiological systems manifest on multiple time scales, and these interactions may differ in healthy versus unhealthy individuals (Ivanov et al. 1999). To assess measures of complexity and interaction over multiple time scales, coarse-grained time series were constructed by averaging the data points within non-overlapping windows of increasing length. The number of time scales τ corresponds to the number of coarse-grainings performed. For the τ^{th} time scale, each element of the coarse-grained time series, $y_j^{(\tau)}$, is given by

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i \tag{5.2}$$

where τ represents the scale factor and $1 \leq j \leq N/\tau$. The first time scale corresponds to the original time series, the second time scale corresponds to one coarse-graining, etc.

5.3.8 Sample entropy

Sample entropy H is a metric of signal complexity described in 2.1. Here we calculated MSE H_τ where τ indicates the number of coarse-grainings performed. Optimal parameter values m, r, and the number of coarse-grainings $\tau_{\max}$ were selected via Bayesian optimization (Ghassemi et al. 2014; Shahriari et al. 2016). The same parameter values were used for calculating sample entropy for both HR and activity, i.e. $m_{\text{HR}} = m_{\text{act}}$, and $r_{\text{HR}} = r_{\text{act}}$.

5.3.9 Mutual information

The mutual information of two discrete random variables X and Y, given by $\mathcal{I}(X;Y)$, measures how much knowing one of the two variables reduces uncertainty about the other. Mutual information was calculated over multiple time scales, e.g. multiscale mutual information (MMI). Significant instances of mutual information were determined by Monte Carlo surrogates, i.e., data randomly shuffled in time. For each subject and time scale, mutual information were computed for 100 surrogates. Transfer entropy from the original source time series was deemed to be statistically significant if it was greater than the 95^{th} percentile of the surrogate results (Kantz et al. 2004).

$$I(X;Y) = \sum \sum p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{5.3}$$

where $p(x,y)$ is the joint probability function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y respectively.

For example, if X and Y are independent, then knowing X does not give any information

about Y and vice versa, so their mutual information is zero.

5.3.10 Darbellay-Vajda (D-V) adaptive partitioning

The computation of transfer entropy and the transformation of time series into a network representation requires estimating joint probability density functions (PDFs). PDFs were estimated via the D-V adaptive partitioning algorithm, in which two time series X and Y are substituted with their ranks ranging from 1 (smallest value) to N (largest value) in sorted X and Y, in a manner similar to some non-parametric statistical tests. The transformed time series of X and Y are $U = \{u_1, u_2, ..., u_N\}$ and $V = \{v_1, v_2, ..., v_N\}$. The two-dimensional space defined by u_{i-t} and v_{i-w} is then recursively partitioned into squares of varying sizes. Initially the space is divided into four equal quadrants where boundaries are at the mid-points. The null hypothesis that data points are evenly distributed across the four quadrants is tested via the χ^2 statistic (Hudson 2006; Lee et al. 2012):

$$s_{\chi^2} = \sum_{i=1}^{4} (M_i - \mu_M^2) \tag{5.4}$$

where M_i is the number of data points in the i^{th} square and μ_M is the mean number of data points per square. If s_{χ^2} is greater than the critical chi-square statistic value for $p = 0.05$, $\chi^2_{95\%}$, with $n^2 - 1$ degrees of freedom where n is the number of dimensions or time series being partitioned, the null hypothesis is rejected, the distribution of the data is not uniform, and the partitioning continues such that the quadrant is split into four sub-quadrants. The partitioning process continues recursively until all partitions satisfy the χ^2 test for containing equal proportions of data. If $s_{\chi^2} \leq \chi^2_{95\%}$, the null hypothesis is not rejected, the partitions in the current iteration are discarded, and the current four quadrants are considered to be one partition. Squares that do not contain data do not contribute to the estimation of transfer entropy or network representations of time series.

5.3.11 Transfer entropy

Transfer entropy given by $\mathcal{T}_{X\to Y}$ is a measure of directional coupling between two concurrently sampled time series $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$, and is defined in more detail by Equation (4.1). Optimal parameter values of t (time lag in X), w (time lag in Y), and $\tau_{\max}$ were selected via Bayesian optimization (Ghassemi et al. 2014; Shahriari et al. 2016).

Significant information flows were determined by Monte Carlo surrogates, i.e., temporally shuffled time series were created and evaluated for larger values of $\mathcal{T}_{X\to Y}$ than the original. For each subject, time scale, and parameter value, transfer entropies were computed for 100 surrogates generated by randomizing the order of HR and activity time series. Transfer entropy from the original source time series was considered statistically significant if it was greater than the 95^{th} percentile of the surrogate results.

5.3.12 Multiscale network representations of time series

Lagged time series over several time scales were converted into multidimensional networks following the methods described by Shashikumar *et al.* (Shashikumar et al. 2017a). A map M from the time series $X \in T$ to a network $g \in G$ can be given by $M : T \Rightarrow G$, where $X = \{X_1, X_2, ..., X_k\}$, k is the total number of time series being considered, and $X_i \in \mathbb{R}^L$, with L being the length of the time series, and $g = \{S, A\}$ consisting of a set of nodes S and adjacency matrix A. The total number of nodes N correspond to the total number of partitions obtained from the D-V partitioning algorithm. Each partition p_i $(i = 1, ...N)$ is assigned to a node $n_i \in N$ in the graph g. Every data point in X is assigned to one of the partitions. The adjacency matrix A is a NxN matrix where a_{ij} corresponds to the transition from node n_i to node n_j. Transitions from node n_i and n_j are represented by the weight a_{ij}. Attributes of this graph, i.e. multiscale network representation (MSNR) features, were used to classify illness. This process was repeated for time series after several coarse-grainings, to construct networks over multiple time scales.

Network attributes were computed using the `Matlab Tools for Network Analysis` open source toolbox from the Strategic Engineering Research Group (Massachusetts Institute of Technology 2011):

- **Number of nodes** are the total number of nodes in the network.
- **Number of edges** are the total number of edges in the network.
- **Link density** is the total number of edges divided by the maximum possible edges in the network.
- **Average degree** is the average value of the degree of all nodes in the network, where the degree of a node is defined as the total number of its neighboring edges.
- **Number of loops** are the total number of independent loops in the network, also know as the "cyclomatic number" or the number of edges that need to be removed so that the network cannot have cycles.
- **Loop3** are the total number of loops of size 3 in the network.
- **Loop4** are the total number of loops of size 4 in the network.
- **Average clustering coefficient** $c(u)$ for node u is defined as the ratio of the number of edges between the neighbors of u to the number of possible edges between them; the average clustering coefficient $C(G)$ of a network is the average of $c(u)$ over all the nodes in the network.
- **Assortativity coefficient** is the Pearson correlation coefficient r of degree between pairs of linked nodes. Positive values of r indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree. In general, r lies between -1 and 1. When $r = 1$, the network is said to have perfect assortative mixing patterns, when $r = 0$ the network is non-assortative, while at $r = -1$ the network is completely disassortative.
- **Algebraic connectivity** is the second smallest Eigenvalue of the Laplacian matrix of a network, where the Laplacian matrix is the difference between the sum of degrees of the diagonal elements in adjacency matrix and the adjacency matrix.

- **Closeness centrality** $cc(u)$ for node u is the inverse of sum of distance from node u to all other nodes in the network, where the closeness centrality of a graph is the average mean of the above is the average of $cc(u)$ taken over all the nodes in the network.
- **Average eccentricity** is determined by first computing the eccentricity of a node or vertex u, defined as $e(u) = max\{d(u, v) : v \in V\}$, where the distance $d(u, v)$ is the length of the shortest path from u to v, and V is the set of all nodes. The average effective eccentricity is the average of effective eccentricities over all nodes or vertices in the network.
- **Maximum effective eccentricity** is also known as the effective diameter, and is defined as the maximum value of effective eccentricity over all nodes in the graph.
- **Trace** is the sum of the eigenvalues of the adjacency matrix, $\sum \lambda$.
- **Energy** is the squared sum of the eigenvalues of the adjacency matrix A, $E(G) = \sum_i^n \lambda_i^2)$.

5.3.13 Binary classification of illness status

Features were used to train a support vector machine (SVM) algorithm with a linear kernel to classify subjects into the schizophrenia or healthy control class, i.e. to perform a binary discrimination task. Classifier performance was assessed via subject-wise leave-one-out cross-fold validation (LOOCV). Given N patients, $N-1$ patients are used to train the classifier and the remaining patient is used as the test set. Features in the training set were transformed to have Gaussian distributions using either the identity, square root, or logarithmic transformations. The transformation resulting in the most normal data, e.g. the lowest k-statistic using the Lilliefors test, was determined from the training set and then applied to the test set to prevent leakage of information. Data in both training and test sets were normalized by subtracting the training mean and dividing by the training standard deviation. Predictions for each subject – defined as the probability of having a diagnosis of schizophrenia – were pooled across crossfolds to report a single pooled area under the receiver operating curve

(AUC; Airola et al. 2009). AUCs were calculated and reported for both training and test sets, and different models (i.e. the set of features used to train the SVM) were compared by calculating the integrated discrimination improvement (IDI; Pencina et al. 2008), given by:

$$\mathrm{IDI} = (\mathrm{S_{new}} - \mathrm{S_{old}}) - (\mathrm{P_{new}} - \mathrm{P_{old}}) \tag{5.5}$$

where S is the integral of sensitivity over all possible cut-off values over $(0, 1)$ interval, P is the integral of $1-$specificity, and "new" and "old" refer to the two models being compared (Pencina et al. 2008).

An asymptotic test for the null hypothesis of IDI $= 0$ was performed, and the P-value reported:

$$z = \frac{\hat{\mathrm{IDI}}}{\sqrt{(\hat{\mathrm{s}}_{\mathrm{events}})^2 + (\hat{\mathrm{s}}_{\mathrm{nonevents}})^2}} \tag{5.6}$$

where $\hat{\mathrm{s}}_{\mathrm{events}}$ and $\hat{\mathrm{s}}_{\mathrm{nonevents}}$ are the standard error of paired differences of new and old model-based predicted probabilities across all event and non-event subjects respectively.

5.4 Results

5.4.1 Mutual information

Mutual information between HR and activity ($\mathcal{I}(HR; \mathrm{act})$) was calculated over four time scales τ_{1-4}. $\mathcal{I}(HR; \mathrm{act})$ for patients with schizophrenia and controls and compared via the two-sided Wilcoxon rank-sum test. $\mathcal{I}(HR; \mathrm{act})$ significantly different for the first three timescales ($P < 0.05$), whereas the difference in medians was not significant for $\tau = 4$ (Figure 5.2A). The opposite trend was observed in the AFib group; for all time scales, patients exhibited significantly lower values of $\mathcal{I}(HR; \mathrm{act})$ compared to controls ($P < 0.05$; Figure 5.2B).

Significance of mutual information between HR and activity was estimated using surrogates whereby each time series was shuffled 100 times and $\mathcal{I}_{surrogate}$ was calculated for each

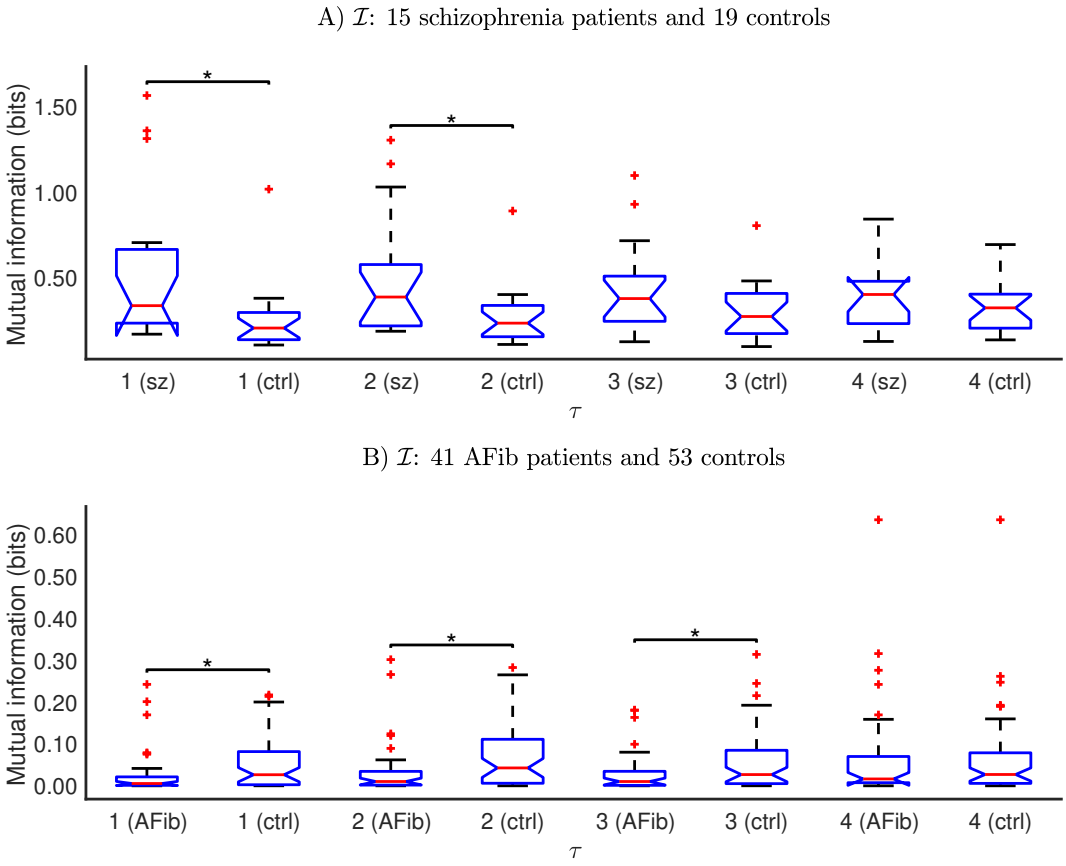


Figure 5.2: MMI between HR and activity for A) patients with schizophrenia and controls, and B) AFib patients and controls. Data is shown via notched box plots; the horizontal red line denotes the median, the notches denote $95th$ percent confidence intervals of the median, borders of the blue box denote the $25th$ and $75th$ percentiles, and the lower and upper whiskers denote the minimum and maximum values, respectively. The y-axis is mutual information in bits, and the x-axis denotes different time scales and sick versus healthy subjects. Asterisks indicates $P<0.05$ via the Wilcoxon rank-sum test.

instance. A mutual information ratio was calculated. The numerator was the 95th percentile of $\mathcal{I}_{surrogate}$ and the denominator was $\mathcal{I}$ of the original data. For all time scales, this ratio was significantly less than the red dashed line of unity for patients with schizophrenia and controls, demonstrating significant mutual information between HR and activity (supplemental Figure A1). In contrast, patients with AFib displayed mutual information ratio metrics near or greater than unity, suggesting the observed values of $\mathcal{I}$ could have been due to random chance. However, for control subjects the mutual information ratio was close to or slightly below unity for all time scales except $\tau = 4$, for which the ratio was slightly above unity.

5.4.2 Transfer entropy

MTE from HR to activity ($\mathcal{T}_{\mathrm{HR}\to\mathrm{act}}$) and from activity to HR ($\mathcal{T}_{\mathrm{act}\to\mathrm{HR}}$) were calculated for patients and controls in both the schizophrenia and AFib groups. In the schizophrenia group, $\mathcal{T}_{\mathrm{HR}\to\mathrm{act}}$ was higher in patients than in controls for the first three time scales ($\tau = 1, 2, 3$), but did not differ for $\tau = 4$ (Figure 5.3A). Similarly, $\mathcal{T}_{\mathrm{act}\to\mathrm{HR}}$ was higher in patients than in controls but for all time scales (Figure 5.3B). In the AFib group, both $\mathcal{T}_{\mathrm{HR}\to\mathrm{act}}$ and $\mathcal{T}_{\mathrm{act}\to\mathrm{HR}}$ were lower in patients with AFib than in controls for all time scales (Figure 5.3C & D).

Following a similar approach as described earlier, a transfer entropy ratio was calculated, with the numerator being 95th percentile of $\mathcal{T}_{\mathrm{HR}\to\mathrm{act,\ surrogates}}$, and the denominator being $\mathcal{T}_{\mathrm{HR}\to\mathrm{act}}$ of the original data (supplemental Figure A2). Patients with schizophrenia had ratios significantly less than unity, suggesting statistically significant values of $\mathcal{T}_{\mathrm{HR}\to\mathrm{act}}$. In contrast, both AFib patients and controls had ratios close to or above 1, suggesting observed non-significant flow of information from HR to activity in that cohort. The flow of information from activity to HR, $\mathcal{T}_{\mathrm{act}\to\mathrm{HR}}$, was assessed the same way, and found to be significant for both patients and controls in both the schizophrenia and AFib cohorts.

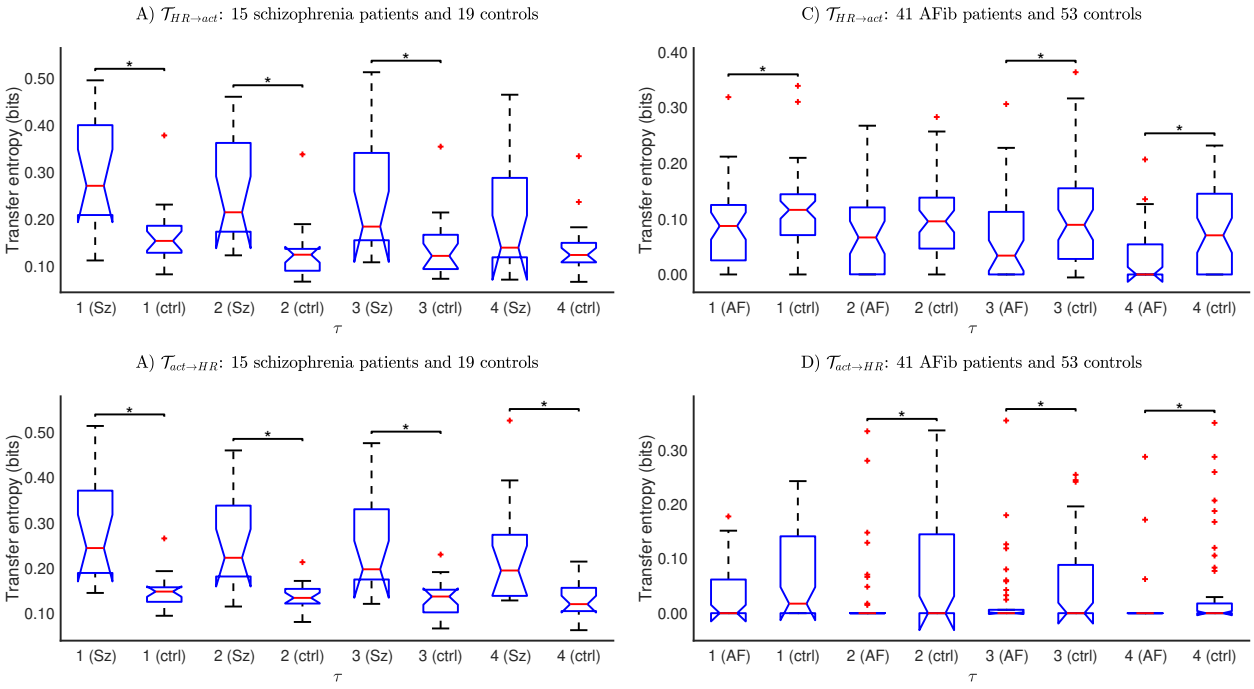


Figure 5.3: MTE from A) HR to activity ($\text{TE}_{\text{HR}\to\text{act}}$) for patients with schizophrenia and controls, B) activity to HR ($\text{TE}_{\text{act}\to\text{HR}}$) for patients with schizophrenia and controls, C) A) HR to activity ($\text{TE}_{\text{HR}\to\text{act}}$) for AFib patients and controls, and D) activity to HR ($\text{TE}_{\text{act}\to\text{HR}}$) for AFib patients and controls. Data is shown via notched box plots; the horizontal red line denotes the median, the notches denote $95th$ percent confidence intervals of the median, borders of the blue box denote the $25th$ and $75th$ percentiles, and the lower and upper whiskers denote the minimum and maximum values, respectively. The y-axis is transfer entropy in bits, and the x-axis denotes different time scales and sick versus healthy subjects. Asterisks indicates $P<0.05$ via the Wilcoxon rank-sum test.

5.4.3 Network representations of time series

MSNR were constructed from HR and locomotor activity time series data. For the schizophrenia group, three time scales were optimal (Figure 5.4A), whereas for the AFib group, the first time scale was optimal (Figure 5.4B). Gross differences in network structure, measured by complexity, node count and edge count, varied both by patient type and time scales.

5.4.4 Classifier performance

Nine feature groups were used to train a support vector machine: 1) statistical moments, 2) MSE, 3) MMI, 4) MTE, 5) MSNR, 6) MSE and MTE, 7) MSE and MSNR, 8) MTE and MSNR, and 9) MTE, MSE, and MSNR. LOOCV was performed to assess classifier performance. Receiver operating characteristic curves (ROCs) were plotted for schizophrenia (Figure 5.5A) and AFib cohorts (Figure 5.5B), and areas under the ROCs (AUCs) were reported for training and test sets (Table 5.1).

In the schizophrenia cohort, MSNR features – alone or in combination with any other feature – resulted in perfect classifier performance in both training and test data (Table 5.1). Model performance was compared using the IDI. MTE demonstrated improved performance versus MMI, and MSNR demonstrated improved performance versus MMI, with $P < 0.05$ for each (Table 5.2).

In the AFib cohort, MSE resulted in the maximum test AUC. MTE alone outperformed MSNR for both training and test AUCs, but did not improve classifier performance when used in combination with MSE. Model performance on test set data was compared using the IDI. MMI outperformed MSE, and MSNR outperformed MMI (Table 5.2).

5.5 Discussion

We assessed interactions between HR and activity by calculating mutual information, transfer entropy, and network representations of time series over multiple time scales. Construct-

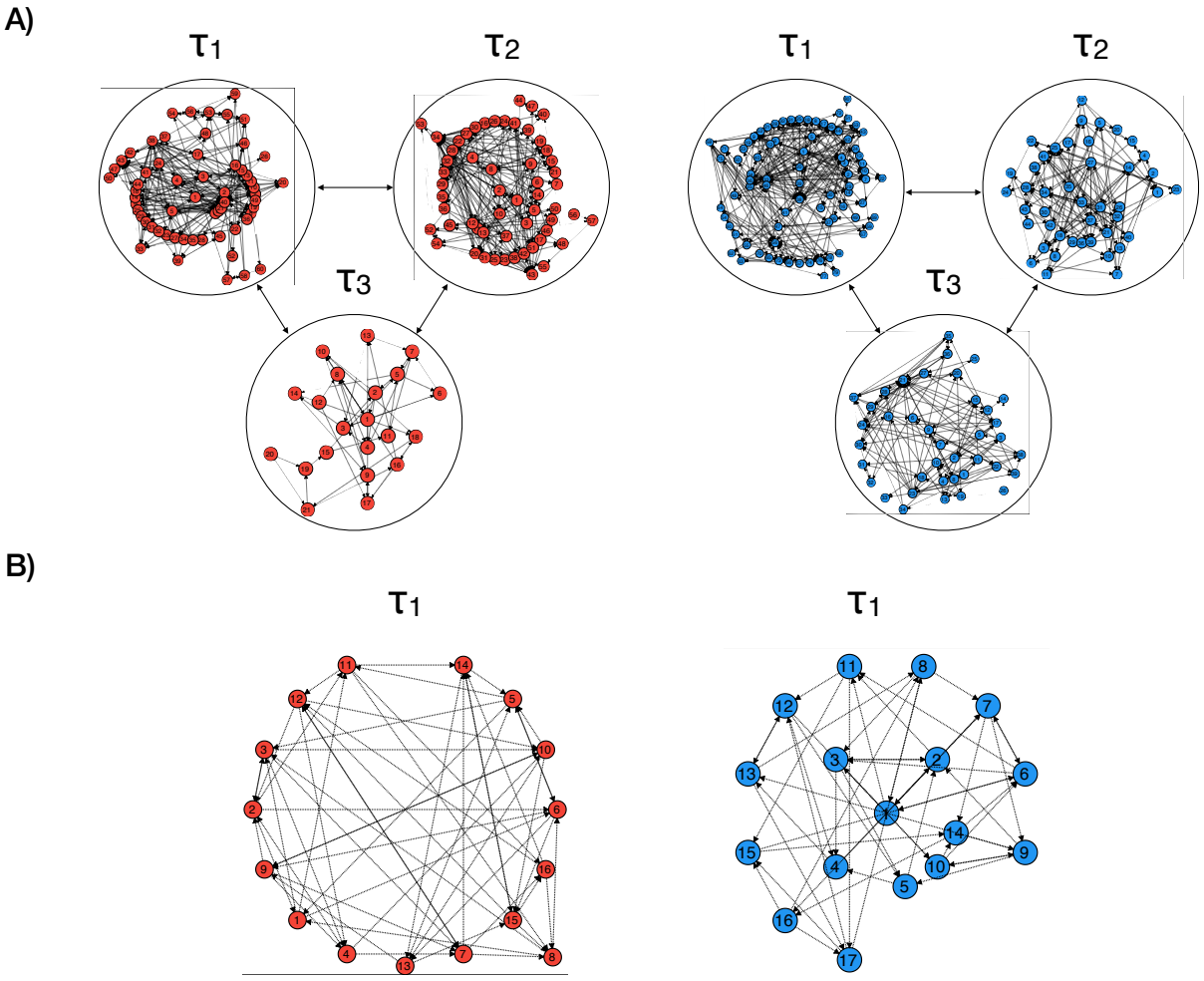


Figure 5.4: MSNR of HR and activity data; each colored circle represents a six-dimensional state defined by a value of HR (e.g. 74 BPM), locomotor activity (e.g. RMS of accelerometry value of 1.7), two time-lagged values of HR, and two time-lagged values of activity. Thus, each state represents a temporal trajectory through physiological and behavioral states. Lines between nodes denote transitions in time from one node to the next. A) Network representations of data from a single subject with schizophrenia (denoted in red) demonstrate a higher number of physiological and behavioral states at τ_2, and a lower number of states at τ_3, compared to states from a healthy control subject (denoted in blue). τ_i indicates the i_{th} time scale. B) Network representations of data from a single subject with AFib (denoted in red) demonstrate a higher number of physiological and behavioral states and more state transitions compared to a healthy control subject (denoted in blue). The properties of these networks were quantified using graph theoretical approaches, and these properties were used as features to train a support vector machine to classify patients from healthy controls.

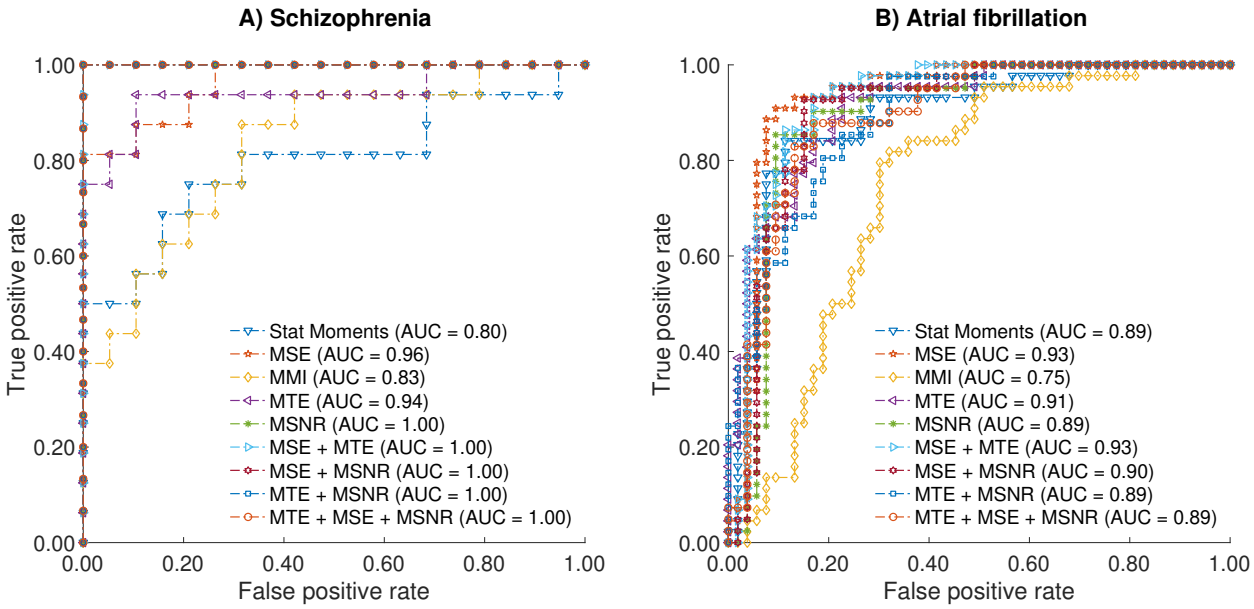


Figure 5.5: ROC curves of models for classifying patients with A) schizophrenia or B) AFib from healthy controls using combinations of different features. Features were calculated from at least ten continous days of HR and locomotor activity. Stat Moments is statistical moments, MSE is multiscale entropy, MMI is multiscale mutual information, and MSNR is multiscale network representations. The Y-axis is the true positive rate and the X-axis is false positive rate.

ing a network from HR and activity time series is a novel approach that utilizes the D-V partitioning algorithm, which is computationally fast and does not require the specification of as many hyperparameters as other partitioning methods (Hudson 2006). Measures of interactions between HR and activity as well as attributes of each signal were calculated and used to train a machine learning algorithm to classify schizophrenia subjects from controls. Perfect classification accuracy was achieved in the schizophrenia cohort using MSNR features, whereas combined univariate analyses on separate HR and activity resulted in lower AUCs. On the other hand, network features did not add significant differentiating power to the classifier when evaluating a non-mental population with AFib. To our knowledge, this is the first use of interactions between HR and activity to distinguish patients from controls.

Univariate features such as statistical moments and MSE were less predictive compared to interaction features, yet enabled classification of schizophrenia significantly better than chance. MSE outperformed statistical moments in both schizophrenia and AFib groups.

Table 5.1: AUCs indicating classifier performance for nine feature groups, or models. The model is described in column 1, results from the schizophrenia cohort are reported in columns 2-3, and results from the AFib cohort are reported in columns 4-5.

	Schizophrenia		Atrial fibrillation	
Model	Training AUC	Test AUC	Training AUC	Test AUC
Statistical moments	0.98	0.84	0.91	0.89
MSE	0.98	0.95	0.94	0.93
MMI	0.90	0.81	0.81	0.77
MTE	1.00	0.94	0.96	0.91
MSNR	1.00	1.00	0.92	0.90
MSE + MTE	1.00	1.00	0.98	0.92
MSE + MSNR	1.00	1.00	0.96	0.90
MTE + MSNR	1.00	1.00	0.99	0.89
MTE + MSE + MSNR	1.00	1.00	0.99	0.89

Measures of complexity of physiological and behavioral time series may better capture mental illness-associated loss in system autoregulation compared to simpler descriptions of the distribution of data (Berle et al. 2010; Montaquila et al. 2015). These results are consistent with our previous work on classifying patients with schizophrenia from healthy controls by learning univariate features of complexity where the test AUC did not reach 1.00 (Reinertsen et al. 2017b).

Mutual information $\mathcal{I}$ quantifies linear and nonlinear dependence between two variables (Duncan 1970). If HR contains information about, or vice-versa, $\mathcal{I}$ will be > 0. $\mathcal{I}$ is likely to be nonzero in the groups studied here because activity can lead to an increase in HR due to a rise in peripheral oxygen demand, and increased HR can precede a rise in activity due to a behavioral response to external cues. We assessed if $\mathcal{I}$ across several time scales contributed to classifier performance. In patients with schizophrenia, MMI resulted in lower classification performance compared to other types of features, albeit better than random chance, with a test AUC of 0.83 (Figure 5.2). In patients with AFib, MMI enabled classification of AFib patients with a test AUC of 0.75.

To assess if mutual information between HR and activity was due to random chance, we shuffled each time series 100 times, calculated $\mathcal{I}_{surrogate}$ each time, and calculated the

Table 5.2: Comparison of model performance on test set data via the IDI. A positive IDI with $P < 0.05$ indicates the new model achieves a statistically significant improvement in classification performance versus the old model. Models are listed in column 1, results from the schizophrenia cohort are reported in columns 2-3, and results from the AFib cohort are reported in columns 4-5.

	Schizophrenia		Atrial fibrillation	
Models	IDI	P-value	IDI	P-value
MSE vs. MMI	-0.316	0.073	-0.297	0.005
MMI vs. MTE	0.277	0.047	0.184	0.064
MMI vs. MSNR	0.449	0.009	0.243	0.026
MSE vs. MTE	-0.034	0.820	-0.113	0.213
MSE vs. MSNR	0.133	0.317	-0.054	0.452
MTE vs. MSNR	0.172	0.215	0.059	0.560

ratio of $\mathcal{I}$ of the original data to the 95th percentile of $\mathcal{I}_{surrogate}$. This ratio was far less than unity for both patients and controls in the schizophrenia group, suggesting significance (supplemental Figure A1). Interestingly, this was not the case for the AFib group; the median ratio was close to and slightly above unity for patients whereas the median was below or close to one for controls. These data suggest a modest reduction in coupling between HR and activity in patients with AFib compared to controls within the same group, but more broadly demonstrate a difference in mutual information by group that may be due to measurement method (Proteus patch for schizophrenia subjects versus SimBand smartwatch for AFib subjects) rather than illness class. SimBand recordings were much shorter at only five minutes long, so there are far fewer gross movements. Longer recordings to capture broader time scales may be necessary to reveal interactions between HR and activity. However, differences between the two groups did not introduce bias, as the classification task was to dichotomize patients from controls within the same group, rather than to distinguish schizophrenia, AFib, and healthy controls in a trinary classification task.

Directed information flow between HR and activity may be more predictive than asymmetric $\mathcal{I}$. Thus we calculated transfer entropy $\mathcal{T}_{HR\to act}$ and $\mathcal{T}_{act\to HR}$. Both $\mathcal{T}_{HR\to act}$ and $\mathcal{T}_{act\to HR}$ were higher in patients with schizophrenia than controls for all time scales (Fig-

ure 5.3). We also assessed significance of transfer entropy using surrogate shuffling. The median ratio of the 95th percentile of $\mathcal{T}_{surrogate}$ to $\mathcal{T}_{HR\rightarrow act,original}$ was less than unity for all time scales and for both patients with schizophrenia and controls, although ratios were slightly lower in patients with schizophrenia versus controls (supplemental Figure A2). Similar to the surrogate analysis of $\mathcal{I}$, this ratio was slightly greater than unity for patients and controls in the AFib group. In contrast, $\mathcal{T}_{act\rightarrow HR}$ was equal to or very close to zero for all subjects in both the schizophrenia and AFib groups, demonstrating significant directed transfer of information from activity to HR regardless of illness group or control status.

MTE outperformed MMI features for both the schizophrenia and AFib group, suggesting the direction of information transfer between HR and activity can be used to distinguish mental illness from health, or cardiac illness from health. Increases in activity following an increase in HR is partially mediated by the ANS, and ANS dysfunction occurs in schizophrenia. $\mathcal{T}_{HR\rightarrow act}$ and $\mathcal{T}_{HR\rightarrow act}$ may be predictive features because of altered coupling between HR and activity in schizophrenia.

Surprisingly, MMI or MTE features enabled classification of AFib with performance superior to random chance. The AFib cohort served as a control; patients did not hvae psychiatric illness, and they were assessed in a seated position in a clinical lab setting that ostensibly reduces the likelihood of significant HR-activity interactions. However, AFib can cause symptoms such as palpitations and dyspnea (Lip et al. 2016; in addition, irregular pulsations conducting through the arm in AFib may result in subtle effects on movement detectable by a wristband. These factors could have contributed to observed associations between HR and locomotor activity in patients with AFib, even when they were asked to sit quietly.

Directed transfer of information between two signals likely occurs at specific regions in time rather than constantly throughout. The distribution of these regions could be a more nuanced and predictive feature of illness than global measures of information transfer. However, such distributions are not captured by mutual information or transfer entropy

of entire time series. To better assess system dynamics and interactions between signals, we constructed network representations of HR and activity time series (Shashikumar et al. 2017a). Each node represents a physiological and behavioral state, and was formed from a partition in a six-dimensional space comprised of lagged forms of HR and activity: 1) HR(t), 2) HR($t-1$), 3) HR($t-2$), 4) act(t), 5) act($t-1$), and 6) act($t-2$). Our approach exploits Takens' theorem, which describes how a dynamical system can be reconstructed from a sequence of lagged observations, given a sufficient embedding dimension $\mathcal{D}$ (Takens 1981). Although the optimal $\mathcal{D}$ is unknown, a lagged embedding approach can yield more information about the properties of a dynamical system compared to analyzing only the observed time series.

Networks over various time scales from a representative subject with schizophrenia and AFib are shown in Figure 5.4. Although the number and connectivity of nodes exhibit variance across subjects within an illness group, networks from patients with schizophrenia visually differed from networks derived from controls. Because the maximum number of edges in a directed graph with n nodes is $n(n-1)$, even slight differences in node count are amplified in other network properties that correlate with edge count, connectivity, and complexity. These data suggest a decreased number of physiological states and transitions in schizophrenia, consistent with previous reports of more structured patterns of cardiac physiology and behavior compared to healthy controls (Chang et al. 2009; Berle et al. 2010; Montaquila et al. 2015).

On the other hand, networks from patients with AFib featured similar numbers of nodes but greater connectivity between nodes compared to controls. Severe AFib is characterized by an "irregularly irregular" heart rhythm, which presumably results in a larger number of physiological states compared to a healthy subject. However, the occurrence of AFib may have been rare in our patient cohort, thus not significantly adding to the number of HR-activity states. Despite a similar number of nodes, transitions between nodes were sufficiently altered in AFib to enable discrimination of patients from controls.

Corroborating the markedly discernible visual differences in networks, MSNR features enabled the classifier to perfectly discriminate patients with schizophrenia from healthy controls, with train and test AUCs of 1.00 (Table 5.1 and Figure 5.5). However, MSNR features did not outperform MTE features for AFib patients. These results indicate a difference in interactions and temporal structure between HR and activity in patients with mental versus cardiovascular illness.

Finally, a comparison of significant differences in classifier performance via IDI in the schizophrenia cohort demonstrated statistically significant improvements of MTE versus MMI, and MSNR versus MMI, with $P < 0.05$ (Table 5.2). In the AFib cohort, MMI versus MSE and MSNR versus MMI were significantly different. MSNR outperformed MTE in the schizophrenia cohort, but MTE outperformed MSNR in the AFib cohort, which may suggest MTE and MSNR capture different mechanisms of coupling between heart rate and movement. Additionally, the probability of a statistically significant difference will increase in a larger cohort, but effect sizes of each type of feature may also differ in mental versus cardiovascular patients, so these results should not be over-interpreted. Moreover, a lower P-value does not necessarily correspond to a superior feature or model (Lo et al. 2015). It should be noted that AUCs from different models can be definitively compared without significance tests when performing LOOCV – a model achieving an AUC of 1.00 perfectly classifies all subjects, and this performance is deterministic and repeatable insofar as the features used to train the model are non-stochastic and do not vary across experiments.

Network representations captured more illness-related information compared to other information theoretical measures of complexity and interaction. Fasmer *et al.* evaluated time series of locomotor activity over 12 days from depressed and schizophrenic patients, mapped these data to a graph, and evaluated measures of complexity (Fasmer et al. 2018). Depressed patients were found to be significantly different from both controls and schizophrenic patients, with evidence of less regularity of the time series. However, the nodes in the graph were univariate and only comprised of motor activity states, whereas the work here reports

multivariate states comprised of both HR and activity. Campanharo *et al.* qualitatively evaluated network representations of chaotic Lorenz and Rossler equations, but the relationship between network attributes and properties seen in physiological data such as noise, autocorrelation, periodicity, and non-stationarity remain unknown (Campanharo et al. 2011). Simple dynamical systems can be generated with known and varying levels of these properties, and attributes of the resulting network representations can be studied. Understanding this mapping could yield insight about the physiological meaning of altered interactions between HR and activity time series in illness.

We note several limitations of this study. Each group was small, consisting of 16 patients with schizophrenia and 19 controls, and 41 AFib patients and 53 controls. We used LOOCV to estimate generalizability of our classifier, but performance metrics may differ for a new group. Comprehensive feature selection was not performed; rather, features from one or several feature groups in combination were compared. Only an SVM classifier was used, but a different machine learning algorithm could achieve better classification performance.

ANS dysfunction in schizophrenia may be partially mediated by inflammation. The concentrations of IL 1-β, IL-6, and transforming growth factor-β have been shown to vary with clinical status, may predict subsequent relapse, and are known to also affect cardiovascular function (Kirkpatrick et al. 2013). Inflammation is inversely associated with changes in time- and frequency-domain HRV measures, even in sub-clinical adult populations without serious cardiovascular disease (Haensel et al. 2008). Regarding behavior, alterations of locomotor activity in schizophrenia have been attributed to cognitive dysmetria, or functional disconnectivity between neural centers of cognition, motor function, and coordination (Honey et al. 2005). However, most studies of schizophrenia did achieve data collection at a sufficiently high frequency to discern more nuanced measures such as directed transfer of information. Further research is needed to elucidate the mechanisms governing interactions between physiology and behavior in schizophrenia and other mental illnesses, and potentially exploit these mechanisms for clinically useful applications such as relapse prediction or remote monitoring

of the efficacy of new therapeutics.

5.6 Conclusion

We demonstrated measures of multiscale interactions – mutual information, transfer entropy, and network properties – between HR and activity enable discrimination of patients with schizophrenia from controls. We repeated this approach using data from patients with AFib, and found network and interaction features did not improve prediction over complexity measures that ignored interactions between HR and activity. To our knowledge this is the first evaluation of interactions between physiological and behavioral states in a mental health population using data measured via objective, affordable, and non-invasive wearable devices.

CHAPTER 6
CONCLUSION

6.1 Validity

The work presented in this thesis explores opportunities to improve algorithmic discrimination of patients with mental illness from healthy control subjects using supervised learning algorithms. Predictive features from HR and/or locomotor activity data were evaluated during specific times (as a proxy for context), over several time scales, and between signals.

6.1.1 Need

One in four people in the world will be affected by mental or neurological disorders, yet only a small fraction of them will receive treatment due to pervasive underdiagnosis, a lack of trained healthcare professionals, stigma, and other reasons (Sayers 2001). Currently, mental illnesses are diagnosed via clinical interview, in which the psychiatrist asks the patient and family members about characteristic symptoms and social and/or occupational dysfunction. However, brain disorders such as schizophrenia can impair insight and hinder the accuracy of self-reporting, challenging both initial diagnosis. PTSD is often underdiagnosed due to social stigma or lack of care access. Finally, even after a diagnosis is established, the nature of mental illness and how psychiatric care is delivered can also result in suboptimal follow-up and monitoring of clinical status.

A growing understanding of ANS dysfunction in neuropsychiatric illness, advances in technology, and a clinical need for objective and reproducible metrics has motivated exploration of noninvasive monitoring of HR, locomotor activity, and other measures (Chang et al. 2009; Liddell et al. 2016; Vancampfort et al. 2017). Importantly, passive monitoring via digital sensors in smartphones and wearables can yield information about a patient's physiology and behavior in the 99% of the time they are not seeing a clinician Asch et al. 2012).

Digital health approaches – whereby physiology and behavior of patients are measured in near real-time and in their native contexts, features are extracted from these data streams, and machine learning or other computational approaches infer clinically useful information – could provide a richer understanding of the day-to-day variability of neuropsychiatric illness, enable assessment of patient status before (rather than after) symptoms reach a level warranting intervention, and reduce biases and inaccuracy intrinsic in subjective questionnaires (Karow et al. 2008; Copeland et al. 2017).

Although promising, the field of "digital mental health" faces numerous technical challenges and areas that remain unexplored. One such topic is the uncertainty of ascribing meaning to aberrations in a time series that may be instead caused by a non-pathological mechanism. Metadata can aid discrimination here by adding contextual information. For example, GPS readings or social media can distinguish elevated entropy of locomotor activity being due to mental illness, or alternatively due to a subject's participation in a social gathering, cultural event, or travel for leisure. Data fusion and robust estimation from noisy data sources has been explored in the setting of electrocardiography for the detection of abnormal rhythms due to cardiovascular disease (Li et al. 2008; Clifford et al. 2012a), and most studies using HR and activity discard days with or subjects who lack sufficient data. However, few digital health studies employ contextual analysis of data, whereby data is emphasized during periods of presumed high signal-to-noise ratio of an illness-induced feature. In particular, evaluating HRV measures during times when a subject is likely to be asleep may reduce motion artifact and increase the ability of a classifier to detect autonomic dysregulation.

A second topic is differing time scales of relevant physiology and behavior differ. For example, a feature calculated from ten minutes of HR and activity data contains information about the ANS or short-term movements. 24-48 hours of data contains information about systems mediated by circadian rhythms and sleep. A week or more of data contains information about social activity, behavior, and lower-frequency physiological dynamics. Few studies of passive monitoring for mental illness account for varying time scales.

A third topic is measures of interaction between signals, such as HR and activity. In theory, information is transferred between cardiovascular physiology and behavior over several time scales. In normal individuals, circadian rhythms mediate the increase in blood pressure and heart rate during the early morning prior to an increase in consciousness, which in turn leads to waking and locomotor activity in the morning. Conversely, the rising of an individual from a chair leads to a rise in blood pressure, heart rate, and sympathetic tone. These responses, interactions, and transitions between physiological and behavioral states vary over time scale, and are partially mediated by the baroreflex and central command, a feed-forward neural mechanism that contributes to motor and cardiovascular function during arousal and exercise (Hall 2010). Since patients with several types of mental illness – such as schizophrenia or PTSD – have ANS dysfunction, physiological and physical responses to changes in HR and locomotor activity may be abnormal (Chang et al. 2009; Montaquila et al. 2015; Alvares et al. 2016).

6.1.2 Overview of contributions

This work demonstrates classification of mental illness using features from HR and accelerometer data is improved by considering information during specific times, over several time scales, and between signals.

In Chapter 1, an introduction is provided and the problems of data collection, time scales, and interactions between data streams are described.

In Chapter 2, past work is surveyed. The clinical epidemiology, pathophysiology, treatment, and ANS alterations are described for schizophrenia and PTSD. Time- and frequency-domain HRV metrics, entropy, and the effects of medication on HRV are reviewed. Rest-activity characteristics of locomotor activity is described. The growing literature on digital sensors for monitoring neuropsychiatric illnesses is surveyed in four parts, organized into smartphones, wearable accelerometers, Holter monitors, and multimodal sensing. The focus of this chapter is on passive monitoring and analyses of HR and locomotor activity, feature

extraction, and classification or regression of clinically relevant outcomes (Table A4).

In Chapter 3, the classification of PTSD from HR time series is reported. To improve the signal-to-noise ratio of HR data, a HR-based window segmentation approach is proposed whereby five 10-minute segments with the lowest median HR were isolated. This work tested the hypothesis that these segments represented quiescent periods, or times when the subject was least likely to be engaged in volitional motor activity and/or most likely to be sleeping or resting. Single-channel ECG data were collected from 23 subjects with current PTSD, and 25 control subjects with no history of PTSD over 24 hours. RR intervals were derived from these data, cleaned, and used to calculate HR and HRV metrics. Features were derived from 1) RR data from these segments, 2) RR data from five randomly selected 10-minute control segments, or 3) all 24 hours of RR data. Classifier performance was assessed via repeated random sub-sampling validation, and area under the receiver operating characteristic curve (AUC) was calculated. A combination of the four most predictive features derived from quiescent segments resulted in a median AUC of 0.86 on out-of-sample test set data. This was significantly higher than the AUC using 24 hours of data (0.72) or random segments (0.67). These results demonstrate the segmentation of HR data into quiescent periods improves the classification of PTSD from HR and HRV measures compared to the use of all collected data (Reinertsen et al. 2017a).

In Chapter 4, the classification of schizophrenia using varying time scales of HR and locomotor data is reported. This work tested the hypothesis that classifier performance and most predictive features varied with time scale. Features from both HR and locomotor activity data were used to train a classifier to distinguish contiguous days of data as belonging to a schizophrenia patient or a healthy control. HR and physical activity was recorded from 12 medicated subjects with schizophrenia and 12 healthy controls. Derived features included statistical moments, rest-activity metrics, transfer entropy, and multiscale fuzzy entropy. The time scale (e.g. window length) of data was varied from two to eight days, and found to affect classifier performance. An analysis window length of eight days resulted in a test set

AUC of 0.96. Reducing the analysis window length to two days only lowered the AUC to 0.91. The type of most predictive features varied with analysis window length. These results demonstrate that time scale, or total length of recorded data, affects classifier performance and most predictive features (Reinertsen et al. 2017b).

In Chapter 5, the classification of schizophrenia by evaluating interactions between HR and locomotor activity is reported. This work tested the hypothesis that information between HR and locomotor activity is altered in mental illness and relatively less altered in cardiovascular illness, and that this information is useful in a machine learning approach to discriminate patients from controls. HR and locomotor activity were recorded via wearable patches in 16 patients with schizophrenia and 19 healthy controls. Measures of signal complexity and interactions were calculated over multiple time scales, including sample entropy, mutual information, and transfer entropy. A classifier was trained on these features to discriminate patients from controls. Additionally, time series were converted into a network with nodes comprised of HR and locomotor activity states, and edges representing state transitions. Graph properties were used as features. To compare against non-psychiatric illness, the same approach was repeated in 41 patients with AFib and 53 controls. Network features enabled perfect discrimination of schizophrenia patients from controls with an AUC of 1.00 for training and test data. Other bivariate measures of interaction achieved lower AUCs (train 0.98, test 0.96), and univariate measures of complexity achieved the lowest performance. Conversely, interaction features did not improve discrimination of AFib patients from controls beyond univariate approaches. This is the first quantitative evaluation of interactions between physiology and behavior in patients with psychiatric illness (*paper submitted*).

6.2 Limitations

This work faces several limitations. The first limitation is the small sample size in each study. The PTSD work only involved 23 subjects with PTSD and 25 controls, the schizophrenia

cohort consisted of 16 patients with schizophrenia and 19 controls, and the AFib cohort was comprised of 41 patients and 53 controls. These sample sizes may result in models of insufficient variance; in a univariate sense the statistical power may not be large enough to detect smaller effect sizes. Since effect size could not be estimated because previous work this thesis builds upon did not evaluate univariate effects, statistical power could not be estimated, and the rigorous evaluation of generalizability versus sample size in a supervised learning framework is well beyond the scope of this work. The notion of adequate sample size tends to be imprecise, but based on the author's experiences, researchers in both the biomedical sciences and machine learning consider sample sizes below one hundred to be "small".

A second limitation of some of these studies is ECG recordings that are no greater than 24 hours per subject. A home-based continuous physiologic monitoring system could potentially evaluate the efficacy of a therapeutic intervention such as medication or cognitive behavioral therapy. However, doing so would require longer monitoring than 24 hours and additional validation studies. An interesting and clinically important topic to study would be if 24-hour measures could predict therapeutic response at far later time points, such as one week or one month after the intervention. Additional work would need to explore how to reduce false positives and prevent alarm fatigue in the setting of longitudinal monitoring.

A third limitation of the supervised learning approach used in this work – either a logistic regression or support vector machine – is model output of probability of belonging to the ill class, which is merely a coarse proxy for illness severity in the form of ANS dysfunction or locomotor abnormalities. This method would estimate a low probability of illness for a subject who is diagnosed with a mental illness yet has atypically low levels of ANS dysfunction. Other aspects of PTSD or schizophrenia symptomatology described in the DSM-V have yet to be evaluated in the context of HRV measures. Furthermore, the generic clinical descriptor of "motor agitation" or "locomotor disturbances" has not been precisely defined in a quantified manner. Inferring the severity of specific manifestations of mental illness

severity could be useful, but doing so would require larger studies with multimodal data including high-resolution ECG recordings, locomotor activity, and clinical questionnaires or visits to provide labeled data.

A fourth limitation (related to the previous) of concern for studies conducted over several days is the day-to-day fluctuation of symptom severity in schizophrenia and PTSD (Kahn et al. 2015). In the schizophrenia study, classifier output varied from day to day (Figure 4.4a), was based on measures of ANS dysfunction and behavior, and may have reflected fluctuations in illness severity. However, detailed information about daily changes in symptoms, e.g. Brief Psychiatric Rating Scale survey data, was unavailable. Such data would be necessary for training a classifier to estimate illness severity accurately, rather than a general probability of belonging to the ill class.

A fifth limitation of this work is controlling for employment status as a proxy for social routine – which relates to activity and restfulness to some extent. However, other potential confounders that could affect HRV or locomotor activity were not explored. Literature suggests potential confounders such as weight, BMI, diet, smoking status, renal function, etc. have a moderate effect on HRV. However, the dominant factors are mental response (Bernardi et al. 2000) and physical movement (Knoepfli-Lenzin et al. 2010). Stress relates to mental state with physiological and behavioral manifestations. Skin conductance, a biomarker for stress mediated by the sympathetic nervous system, has been shown to differ in schizophrenia patients (Bär et al. 2008). Additionally, cortisol secretion and stress sensitivity may be associated with schizophrenia, or subsequent development of schizophrenia following the prodromal phase of the illness (Walker et al. 2013; Holtzman et al. 2013). Stress affects activity as well as other aspects of mobile device usage; Sano et al. 2013 reported using using screen on, mobility, call or activity level information to distinguish stressed from non-stressed individuals with an accuracy of 75%. Aside from matching employment status, stress was not controlled for in these studies. Doing so with conscious patients is challenging especially in an ambulatory setting. Additionally, stress invariably accompanies social risk

factors such as physical abuse, sexual abuse, maltreatment and bullying, which are associated with increased risk of later schizophrenia (Stilo et al. 2010).

A sixth limitation is how antipsychotic medications affect the illness of interest, thus modifying HRV measurements, but also may exert a direct effect on the cardiovascular system and contribute to ANS dysfunction (Rechlin et al. 1994; Birkhofer et al. 2013; Huang et al. 2013). Mondelli et al. demonstrated that antipsychotic medication can reduce cortisol secretion and normalize HPA-axis hyperactivity in psychotic patients (Mondelli et al. 2010). The literature has conflicting reports; Bär et al. 2008 did not find significant changes in ANS function after antipsychotics were administered to patients, while Henry et al. 2010 et al. found that risperidone, valproate, or mood stabilizers did not significantly affect HRV in bipolar or schizophrenia patients. Our database lacked more detailed information about the type, dose, or adherence of antipsychotic medications taken by patients in the schizophrenia cohort; results herein may not generalize to a non-medicated patient population. However, a classifier affected by a patient's medication could potentially be used to monitor adherence and treatment efficacy.

A seventh limitation is the selection of hyperparameter (e.g. entropy template length and minimum amount of data per day) values from prior studies instead of optimized in a first-principles or data-driven manner. Classifier performance could likely be improved using techniques such as Bayesian optimization (Ghassemi et al. 2014; Shahriari et al. 2016) which were employed in the final chapter of work on interactions between HR and locomotor activity.

An eighth limitation is the use of LOOCV, which has low variance but high bias and computational cost compared to k-fold cross validation methods with fewer folds. The limitation here is not necessarily intrinsic to this method; rather, the issue lies in that only one cross-validation method, and only one (regularized logistic regression or support vector machine) learning algorithm were used for each experiment. A more robust approach would involved comparing several well-established learning algorithms including logistic regression,

support vector machine, random forest, and deep neural networks. Additionally, a larger sample size and the use of a held-out validation set in addition to training and test sets, or an entirely independent and newly collected validation cohort of subjects, would improve the probability that models are capturing a consistent attribute of pathology rather than a unique aspect of the data, and would generalize to a new population.

6.3 Future work

Passive monitoring of physiology and behavior, feature extraction from time series data, and machine learning to estimate, classify, or predict outcomes has the potential to shift today's model of encounter-driven patient care towards ambulatory monitoring and remote management. Several technical topics relevant to this thesis are being actively investigated, including change point detection (CPD), entropy measures, and network dynamics to assess interactions between multivariate data streams. Other areas of future direction and opportunity for the field of mobile mental health include overcoming limitations of clinical trials, addressing challenges of low-resource settings, and using monitoring to affect patient outcomes.

6.3.1 Change point detection

HR and locomotor activity are non-stationarity phenomena – stochastic processes whose unconditional joint probability distribution changes over time (Manuca et al. 1996). Consequently, parameters such as mean and variance also change. In addition to non-stationarity, these time series demonstrate long-range correlations, autoregression, and complex interactions with other systems (Ivanov et al. 1999; Hausdorff et al. 1995; Pedro et al. 2001; Xiong et al. 2017).

Time varying autoregressive and point process based techniques have been developed to quantify the relationship between multiple variables in non-stationarity physiological time series and to extract spectral indices of autonomic control (Barbieri et al. 2008; Gederi et

al. 2014). However, these techniques model each time series individually and it is unclear they can be used to identify dynamic behaviors that could serve as physiomarkers for illness classification. Furthermore, they make the assumption that data is non-stationarity, so segmentation or other methods of removing non-stationarity is required.

One simple approach to reduce nonstationarity and noise from extraneous factors is to evaluate data solely during periods of low activity. This approach whereby quiescent periods of data with lowest median HR were analyzed in lieu of the entire time series was demonstrated to improve classifier performance in a cohort of subjects with PTSD and healthy controls, suggesting restful periods contain features that are more attributable to illness class (Reinertsen et al. 2017a). More sophisticated change point detection approaches could potentially sort data into parametrically similar segments, and even further increase the signal-to-noise ratio of features derived from data within each segment (Adams et al. 2007). A popular techniques to artificially remove non-stationarities is 'detrending' via removing a mean, slope, or nonlinear fit in an arbitrary piecewise manner Lan et al. 2010; Wu et al. 2007. However, detrending tends to create large artifacts around changes in stationarity Raffalovich 1994; Nelson et al. 1981.

Changepoint detection (CPD) – the estimation of points in time where the probability distribution of a stochastic process changes – can enable the analysis of stationary segments of data and reveal underlying structure. This results in the division of time series into segments that meet the criteria for stationarity. Bernaola-Galvàn Pedro et al. 2001 used time series segmentation and CPD to investigate non-stationarities in human HR time series and found mean level jumps between HR segments were smaller in heart failure patients compared to healthy controls. Furthermore, HR interval segments were found to follow a power law distribution, for both heart failure patients and healthy controls.

A variety of CPD methods exist and have been applied to physiological data such as HR, yet little work has been done to determine the optimal algorithm for a given dataset based on its properties or ultimate downstream task (e.g. supervised learning to dichotomize dis-

ease). Cakmak et al. 2018 demonstrated a principled approach for selecting a CPD algorithm for a specific task, such as disease classification. Eight key algorithms were compared, and the performance of each algorithm was evaluated as a function of temporal tolerance, noise, and abnormal conduction (ectopy) on realistic artificial cardiovascular time series data. Artificial data was used because ground truth of change points were known and thus CPD algorithm performance could be assessed, and optimal parameters estimated. On artificial data, Modified Bayesian Changepoint Detection achieved superior positive predictive value (PPV) for identification of change points while Recursive Mean Difference Maximization (RMDM) achieved the highest true positive rate (TPR).

Algorithms were also applied to HR time series data from 22 patients with REM-behavior disorder (RBD) and 15 healthy controls, using the parameters selected using artificial data. Features were derived from the detected changepoints to discriminate patients from healthy controls using a K-Nearest Neighbors approach. Segment lengths – time between estimated changepoints – were fitted to a Pareto distribution and characterized by scale and shape parameters. These two parameters were calculated for each subject and used as features. For classification, a K-Nearest Neighbors (KNN) approach with ten-fold cross validation was used. To find the optimal number of neighbors and distance metric for KNN, Bayesian optimization was performed. The objective function of KNN was defined as the percentage of neighbors belonging to the same class for each point, and the highest area under the precision-recall curve was calculated using this metric. KNN was chosen as a simple example to illustrate the technique, rather than as an optimal classifier. Performance metrics calculated were accuracy, AUC, AUCPR, TPR, PPV, and F1 score. For the classification task, features derived from the RMDM algorithm provided the highest leave one out cross validated accuracy of 0.89 and true positive rate of 0.87.

Automatically detected changepoints provide useful information about subject's physiological state which cannot be directly observed. However, the choice of CPD algorithm depends on the nature of the underlying data and the downstream application, such as a

classification task. This work is the first meaningful comparison of CPD algorithms, and a novel utilization of the technique towards a classification task. Future work in CPD involves additional characterization of methods such as robustness in the presence of noise, utility when combined with features from other data streams, and . CPD applied to HR, activity, and other physiological time series or even questionnaire scores could add predictive value to a supervised learning approach in a variety of illnesses, including neuropsychiatric conditions.

6.3.2 Entropy measures

Entropy is defined as the average amount of information produced by a stochastic source of data. Complex dynamical systems tend to generate time series with high entropy, whereas simple functions such as white noise or sine waves produce time series with low entropy. Entropy of HR or locomotor activity reflect average complexity of the underlying physiological system. These measures, especially over multiple time scales, have been shown to correlate with and even precede clinical illness such as AFib (Shashikumar et al. 2017b), sepsis (Shashikumar et al. 2017a), heart failure (Costa et al. 2002; Liu et al. 2013; Zhao et al. 2015), and mental illness such as schizophrenia and PTSD (Osipov et al. 2015; Reinertsen et al. 2017a).

Sample entropy or "SampEn" (eq 2.1) is perhaps the most popular in the literature. However, SampEn can change significantly and/or non-monotonically with small changes in parameter values and thus exhibits poor statistical stability. Additionally, SampEn only accounts for similar patterns with similar amplitudes, not similar patterns with different amplitudes. These shortcomings have been addressed by replacing the binary Heaviside classifier with a continuous membership degree between 0 and 1, based on fuzzy set theory (Chen et al. 2007), and replacing probability estimation with density estimation for entropy approximation (Lee et al. 2001; Liu et al. 2013; Li et al. 2013). This novel "fuzzy approximation of entropy" improves robustness to noise and classifier accuracy when using entropy

as a feature for machine learning.

Various entropy approaches have been compared for discriminating AF, including fuzzy entropy, sample entropy, coefficient of sample entropy, and a novel extension of fuzzy entropy that adjusts for heart rate by subtracting the natural log value of the mean RR interval (Liu et al. 2018). For classifying AF and non-AF rhythms, fuzzy entropy achieved AUCs of 92.72%, 95.27% and 96.76% for 12-, 30- and 60-beat window lengths respectively. This was higher than the performance of the other techniques.

Entropy calculations require selecting hyperparameters such as the template length m, the radius or normalized distance r within which two sequences are considered a match, and the overall segment length of data to be assessed. Most studies use hyperparameter values from previous literature, rather than a rigorous approach based on first principles. For example, Zhao et al. 2015 recommends $r = 0.1$ times the standard deviation of RR time series, but this observation is empirical and based on results within a single study.

Optimization methods have been explored for parameter estimation, including grid search, random search, and Bayesian optimization. This method treats the unknown objective function as random, and places a prior over it using a Gaussian process (Shahriari et al. 2016). After gathering the function evaluations, which are treated as data, the prior is updated to form the posterior distribution over the objective function. The posterior distribution, in turn, is used to determine what the next query point should be. Bayesian optimization has been utilized for estimating values of sample entropy (Shashikumar et al. 2017a), transfer entropy (Reinertsen et al. 2018), and other information theoretical properties of time series that are used as features for machine learning-based classification of disease.

Areas of future work include other optimization methods such as genetic algorithms. Computational speed is still a challenge in the assessment of long time series. Physiological underpinnings of altered complexity are still being explored both in theoretical and experimental work. Lastly, the analysis of local versus global complexity of time series may enable a dynamic approach whereby the optimal parameters used within a local segment of data

is determined using prior information rather than using a single set of parameters for the entire time series. Such an approach would be complemented by change point detection to select segments, and multiscale analysis to parse complexity over various timescales.

6.3.3 Network dynamics

Network or graph representations of time series may capture different illness-related information compared to other information theoretical measures of complexity of the original time series. Fasmer et al. evaluated time series of locomotor activity over 12 days from depressed and schizophrenic patients, mapped these data to a graph, and evaluated measures of complexity (Fasmer et al. 2018). Depressed patients were found to be significantly different from both controls and schizophrenic patients, with evidence of less regularity of the time series.

Campanharo et al. qualitatively evaluated network representations of chaotic Lorenz and Rossler equations, but the relationship between network attributes and properties seen in physiological data such as noise, auto-correlation, periodicity, and non-stationarity remain unknown (Campanharo et al. 2011). To study these properties, simple dynamical systems can be generated with known and varying levels of these properties and converted into networks. Understanding this mapping could yield insight about the physiological meaning of altered interactions between HR and activity time series in illness.

Multivariate graphical representations can also be used to assess interactions between variables. Prior work on graphical representations of time series have been univariate, comprised only of motor activity states, whereas the work in this thesis and by Shashikumar et al. 2017a explores multivariate states comprised of HR and activity, or HR and BP. Future work should incorporate even higher dimensionality by including other clinically relevant time series data. Early exploration of this multivariate approach is occurring in the intensive care unit, where several types of data are monitored in near-real time at a high sampling frequency. However, network analysis could also be useful in mental health. One example is using natural language processing on social media or passive audio recording to generate a

time series of sentiment, and creating multivariate states or nodes of a graph represented by that sentiment as well as physiological and locomotor activity time series. Finally, data with different formats could be fused into a classifier separately from a network representation of time series. For example, results of lab tests such as serum creatinine represent important clinical information, but are not time series, and could thus be input as features alongside network properties to a machine learning algorithm.

6.3.4 Overcoming limitations of clinical trials

Strict inclusion and exclusion criteria are employed to test interventions against a clean background, rather than a real-world scenario in which adherence to the intervention or data collection protocol can be more challenging. Data are collected from patients using long, paper-based questionnaires, journals, or web-based surveys. These tools are inconvenient and time-consuming to patients and do not reflect the context of their daily lives. Only 2% of the eligible population in the U.S. participate in clinical trials (Woodcock et al. 2017). Those who do participate attend an average of 11 trial site visits over six months which can require traveling a significant distance. Finally, conducting trials for patients with serious neuropsychiatric illness can be especially challenging due to limited ability to adhere to study protocols.

Mobile and internet-connected technologies can help address some of these issues by enabling trials to be carried out at a participant's home or local physician's office – a "virtual" or "remote" trial – rather than at a central trial site (Seyfert-Margolis 2018). Virtual trials could also increase the rate of enrollment in exploratory or clinical studies (Savage 2015). For example, over eight months the MyHeart Counts app attracted over 48,000 people who consented to participate in a study of cardiovascular health; 40,000 people uploaded data including surveys on diet, well-being, risk perception, work-related and leisure-time physical activity, sleep, and cardiovascular health (McConnell et al. 2017). During the initial seven-day monitoring period, participants' motion was recorded via phone accelerometry. After

one week, 4,990 people completed a six-minute walk test. Similarly, the mPower app, built using Apple's ResearchKit framework in a collaboration with the University of Rochester and Sage Bionetworks, aims to quantitatively assess symptoms of Parkinson's disease, and has been downloaded by 48,000 people with 9,520 subjects consenting to sharing their data (Bot et al. 2016). Novartis has worked with Science 37, a technology company that develops decentralized clinical trial technology and design, on virtual trials for cluster headache, acne and nonalcoholic steatohepatitis (NASH) (Novartis 2018). Recently, these two entities announced a strategic alliance to initiate up to 10 new decentralized and technology-driven remote clinical trials over the next three years. In addition to bolstering trial enrollment and retention, digital sensors could detect more subtle or nuanced effects of an intervention that could be missed by traditional outcome measures. The quantity and intrinsic speed of data gathering and processing afforded by sensors and software could also better enable adaptive trials, whereby investigators use accumulated data and modify or redesign the trial while the study is still ongoing (Chow 2014).

Digital sensor data is likely to complement rather than replace data obtained in current research trials such as blood biomarkers and imaging. The Emory Healthy Aging Study is an example of this multifaceted approach and will be the largest clinical research study ever conducted in Atlanta, GA (Emory University 2016). The goal is to develop a midlife biomarker of Alzheimer's disease, since it is now well established that the disease begins about two decades prior to the onset of symptoms. Developing new ways to detect the disease in the asymptomatic phase is key for developing preventative treatments. To accomplish this goal, the Emory Healthy Aging Study first aims to recruit 100,000 individuals to participate in an online study to assess risk factors identified via health questionnaires, smartphones, and wearable devices. The second aim is to deeply phenotype a subpopulation of 3000 of these subjects every few years to assess risk factors by profiling genetics, cardiovascular physiology, blood and spinal fluid biomarkers, and brain and retinal imaging. Analyses of subjects' profiles, including their amyloid status, will facilitate discovery of new biomarkers

with diagnostic and prognostic utility.

6.3.5 Addressing needs in low-resource settings

Mobile and wearable technologies have become dramatically cheaper over the past few decades, and could help address the under-distribution of medications and personnel related to neuropsychiatric care in low-resource settings (Collins et al. 2011). Young males, ethnic minorities and people living in socioeconomically disadvantaged areas are more likely to experience "severe mental disorders including schizophrenia, bipolar disorder, and depression with psychotic symptoms" (Jongsma et al. 2018). Furthermore, even in a wealthy country such as the USA, ethnic minorities have significantly less access to care than do European Americans (Mcguire et al. 2008). Compounding this issue, the poorest countries spend the lowest percentages of their overall health budgets on mental health, and have less relative availability of diagnostic encounters and interventions (Saxena et al. 2007). Telepsychiatry and teleneurology can extend the geographic reach of clinicians in regions with limited health resources, but this approach is still limited by the supply of trained professionals. To deliver interventions in a more scalable manner, smartphone and internet-based methods have been explored, including prerecorded video tutorials, self-help interventions, online communities or support groups, and guides to help patients navigate their healthcare system (Kazdin et al. 2013). Digital sensors could complement these approaches by enabling detection of early signs of illness relapse, medication adherence, or treatment efficacy. Although technology-based care delivery methods such as telemedicine are becoming increasingly available in health systems, passive monitoring has yet to become an established component of clinical workflow, especially in resource-poor regions. Many attempts at delivering affordable healthcare technologies into such environments have not achieved the intended levels of impact due to a focus only on cost or simplicity. Attention to sustainable business practices, local cultural dynamics, and integration with existing resource and workflow may enable the potential of these technologies to educate and assist patients and providers; Clifford

2016 provides a thorough review of these considerations and proposes structural ecosystem changes to help achieve empowerment.

6.3.6 Using monitoring to affect patient outcomes

Although much work has focused on demonstrating feasibility of passive sensing, the gap between data capture and meaningful improvements in patient outcomes has yet to be closed (Patel et al. 2015). A growing body of literature has shown that smartphones not only can monitor patients but can also send information to patients in a way that affects clinical outcomes. SMS can increase adherence to antiretroviral therapy and smoking cessation (Free et al. 2013), and smartphone delivery of cognitive behavioral therapy can reduce anxiety, depression, stress, and substance use (Ehrenreich et al. 2017). Recently, Freeman et al. conducted the largest RCT of a psychological intervention for a mental health problem (Freeman et al. 2017). 3,755 students with insomnia from 26 UK universities were enrolled in the trial, with 1,891 receiving digital CBT for insomnia ("Sleepio"), and 1,864 receiving standard practice treatment. Digital CBT was accessible via web browser, and sleep diaries and relaxation audio was accessible via smartphone. At ten weeks, Sleepio significantly reduced insomnia, paranoia, and hallucinations compared to the usual practice. However, no large RCT focused on neuropsychiatric illness has reported a positive impact of passive monitoring on outcomes. A recent difference-in-differences random effects meta-analysis of RCTs of remote patient monitoring did not find statistically significant impacts on any of six outcomes including body mass index, weight, waist circumference, body fat percentage, systolic blood pressure, and diastolic blood pressure (Noah et al. 2017). Interventions based on health behavior models and personalized coaching – relevant to neuropsychiatric care – were most successful.

6.3.7 Ongoing and future studies of note

Large-scale ongoing and future studies of neuropsychiatric disease are increasing utilizing smart devices as a complementary or even central method of gathering participant data (Table A5). Some studies are described briefly here.

Faurholt-Jepsen et al. will assess up to 400 patients with bipolar disorder, randomizing them to either 1) a smartphone-based monitoring system including a feedback loop between patients and clinicians, and cognitive behavioral therapy, or 2) standard treatment. The study will evaluate outcomes such as re-admissions, symptomatic severity, and quality of life.

Verily, University of North Carolina, and Harvard University are leading the AURORA study, which is a 19-institution five-year endeavor to perform the most comprehensive observational study of trauma to date.

Emory University is conducting the Healthy Aging Study on 100,000 participants. The overarching goal is to develop a midlife biomarker of Alzheimer's disease, since it is now well established that the disease begins about 2 decades prior to the onset of clinical symptoms.

The UCLA Depression Grand Challenge Study aims to enroll 100,000 people in a 10-year study. The aim is to screen for depression, analyze participants' genetics, measure early adversity and life stress and assess symptoms.

These and other studies of similar scope are of interest because they seek to 1) gather data from multiple clinical sites in an effort to demonstrate generalizability of monitoring approaches, 2) link physiomarkers gathered from smart devices with traditional biomarkers from cerebrospinal fluid, blood, genome studies, and imaging, and 3) generate novel hypotheses about disease mechanism, progression, and possible interventional targets.

6.3.8 Closing remarks

This thesis contributes to the effort of illness classification via features from heart rate and accelerometer time series data. Classification performance is improved by considering

information during specific times, over several time scales, and between signals. Passive sensing is an important but early step in the iterative process by which data is used to improve patient management, e.g. revise parameters of CBT or other psychotherapy, adjust doses or selection of pharmacological agents, or modify recommended lifestyle and behavior changes. In turn, the effect of these interventions can be measured close to real-time. Thus, digital sensors will likely form an integral component of how healthcare is delivered in many clinical specialties: a feedback loop starting with data-driven insight about pathophysiology and/or treatment, that in turn optimizes therapy, and ultimately improves patient outcomes.

APPENDICES

A.1. Questionnaires, surveys, and scales

The self-reporting of symptoms is an extremely useful gauge of patient progress or acuity. Although such surveys have been traditionally administered via paper, or more recently via web pages, it is increasingly common to capture such data through an approach called Ecological Momentary Assessment (EMA), whereby questions can be delivered to the subject via smartphone in response to triggers, a certain time, or a pattern of interest in gathered data. The questions can be repeatedly administered if the user does not answer. While there is little evidence so far as to the effect this has on such scales, the flexibility this offers provides a new avenue for research into such systems, whereby timing of the response, and even corrections during the process could be analyzed to extract further information about the state of a patient. In this section we review a variety of the most relevant surveys for neuropsychiatric EMA and provide the evidence base for their traditional use.

The Perceived Stress Scale (PSS) was developed to measure psychological stress, defined as "the extent to which persons perceive that their life demands exceed their ability to cope" (Cohen et al. 1983). The PSS predicts both objective biological markers of aging (Espel et al. 2004), cortisol levels (Malarkey et al. 1995), immune markers (Maes et al. 1999), depression (Carpenter et al. 2004), and increased risk for disease among persons with higher perceived stress levels.

The Hamilton Rating Scale for Depression (HRSD, HAMD, or HAM-D) is a multiple item questionnaire used to quantify the results of an interview assessment of symptoms in an adult patient diagnosed with depression (Hamilton 1960). Severity of depression is assessed by probing mood, feelings of guilt, suicide ideation, insomnia, agitation or retardation, anxiety, weight loss, and somatic symptoms among 17 to 29 dimensions (depending on version; often referred to as the HAMD-17 or HAMD-29 respectively) with a score on a 3 or

5 point scale. A score of 0-7 is considered to be normal. Scores of 20 or higher indicate moderate, severe, or very severe depression, and are usually required for entry into a clinical trial. However, the HRSD has been criticized as a test because it places more emphasis on insomnia than on suicide ideas and gestures (Bagby et al. 2017). An antidepressant may show statistical efficacy even when thoughts of suicide increase but sleep is improved. Alternatively, even if a medication effectively reduces depressive symptoms, if sexual and gastrointestinal symptoms worsen as a side effect, efficacy can be underestimated. Results of a large meta-analysis suggest that HRSD achieves good overall levels of internal consistency, inter-rater and test–retest reliability, but some HRSD items (e.g., "loss of insight") are not sufficiently reliable (Trajković et al. 2011).

The Quick Inventory of Depressive Symptomatology (QIDS-SR16) is a shortened 16-item version of the 30-item Inventory of Depressive Symptomatology (IDS), a structured interview that was constructed by selecting only items that assessed DSM-IV diagnostic criterion items for MDD (Rush et al. 2000). The research group that developed the IDS obtained feedback/critique from more than a dozen, largely US, clinical researchers who were experts in depression. The nine domains of the QIDS-SR16 comprise sad mood, concentration, self-criticism, suicidal ideation, interest, energy/fatigue, sleep disturbance (initial, middle, and late insomnia or hypersomnia), decrease/increase in appetite/weight, and psychomotor agitation/retardation. The total score ranges from 0 to 27. QIDS-SR16 has high internal consistency, as well as high correlation with the IDS and the HAMD (Rush et al. 2003).

The Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire (PHQ) was designed by the PHQ Primary Care Study Group to be a fully self-administered survey; the original survey it was based upon was clinician-administered (Spitzer et al. 1999). There is an optional fourth page that includes questions about menstruation, pregnancy and child-birth, and recent psychosocial stressors. The original PHQ assessed 18 current mental disorders. By grouping several mood, anxiety, and somatoform categories together, the PHQ greatly simplifies the differential diagnosis by assessing only eight

disorders: MDD, panic disorder, other anxiety disorder, bulimia nervosa, other depressive disorder, probable alcohol abuse or dependence, and somatoform and binge eating disorders. Patients indicate for each of the 9 depressive symptoms whether, during the previous 2 weeks, the symptom has bothered them "not at all," "several days," "more than half the days," or "nearly every day.". Patients also indicate for each of the 13 physical symptoms whether, during the previous month, they have been "not bothered," "bothered a little," or "bothered a lot" by the symptom. The PHQ Primary Care Study Group found agreement between PHQ diagnoses and those of independent mental health professionals; for the diagnosis of any 1 or more PHQ disorder, $\kappa = 0.65$; overall accuracy, 85%; sensitivity, 75%; specificity, 90%, similar to the original PRIME-MD questionnaire. Furthermore, in addition to making criteria-based diagnoses of depressive disorders, the PHQ-9 has also been show to be a reliable and valid measure of depression severity (Kroenke et al. 2001). A slightly shorter eight-question version of this survey, the PHQ-8, is also sometimes used.

The Center for Epidemiological Studies Depression (CESD) scale is a short self-report scale comprised of 20 questions that ask how often over the past week a person experienced symptoms associated with depression, such as restless sleep, poor appetite, or feeling lonely (Radloff 1977). Each item is scored 0 to 3: 0 = Rarely or None of the Time, 1 = Some or Little of the Time, 2 = Moderately or Much of the time, 3 = Most or Almost All the Time. Total scores range from 0 to 60, with higher scores indicating greater depressive symptoms. Cutoff scores identify individuals at risk for clinical depression with good sensitivity and specificity, and high internal consistency (Lewinsohn et al. 1997).

The Beck Depression Inventory (BDI) consists of 21 multiple-choice questions that ask how the subject has been feeling in the last week, and is a proxy for a structured clinical interview (Beck et al. 1961). Questions inquire about symptoms of depression such as hopelessness and irritability, cognitions such as guilt or feelings of being punished, physical symptoms such as fatigue, weight loss, and lack of interest in sex. Each question has a set of at least four possible responses, ranging in intensity. A value of 0 to 3 is assigned

for each answer, and the values are summed to calculate a total sum up to 63. A higher total score indicates more severe depressive symptoms. The BDI is one of the most widely used psychometric tests for measuring the severity of depression; its successor is the BDI-II which is now more common. The BDI was revised 1996 to the BDI-II in response to the American Psychiatric Association's publication of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, which changed many of the diagnostic criteria for MDD (Beck et al. 1996). The BDI-II is used to evaluate how the subject has been feeling over the past two weeks instead of one week, in order to be consistent with the DSM-IV time period for the assessment of MDD.

The Young Mania Rating Scale (YMRS) is an eleven-item clinician-administered scale to rate manic symptoms. This score correlated with the number of days of subsequent stay in hospital, and significantly differed in patients before versus after two weeks of treatment (Young et al. 1978). A parent report version of the YMRS (P-YMRS) was assessed in a cohort of 117 youths age 5-17 (Gracious et al. 2002). The P-YMRS demonstrated acceptable internal consistency. Logistic regressions discriminated bipolar mood disorder versus unipolar disorder, versus disruptive behavior disorder, and versus any other diagnosis. Classification rates exceeded 78%, and receiver operating characteristics analyses showed areas under the curve greater than 0.82.

The Altman Self-Rating Mania scale (ASRM) is a self-administered survey that was originally evaluated on a cohort of 22 schizophrenic, 13 schizoaffective, 36 depressed, and 34 manic patients (Altman et al. 1997). The Clinician Administered Rating Scale for Mania (CARS-M) and Mania Rating Scale (MRS) were completed at the same time to measure concurrent validity. Principal component analysis of ASRM items revealed three factors: mania, psychotic symptoms, and irritability. Baseline mania subscale scores were significantly higher for manic patients compared to all other diagnostic groups. Posttreatment scores were significantly decreased in manic patients for all three subscales. ASRM mania subscale scores significantly correlated with MRS total scores ($r = 0.72$) and CARS-M mania

subscale scores ($r = 0.77$). Test-retest reliability for the ASRM was significant for all three subscales. Mania subscale scores of greater than 5 on the ASRM resulted in sensitivity of 85.5% and a specificity of 87.3%.

The General Behavior Inventory (GBI) is a 73-question self-administered survey that evaluates various aspects of mood and is designed to identify the presence and severity of manic and depressive moods in adults (Depue et al. 1981). It consists of two scales to assess depressive symptoms (46 items) and hypomanic / biphasic (mixed) symptoms (28 items) (Youngstrom et al. 2008). GBI items use a Likert scale from 0-3: 0 (never or hardly ever present), 1 (sometimes present), 2 (often present), and 3 (very often or almost constantly present). The GBI has high internal consistency and retest reliability because of its large number of items. Retest reliability also is good over a week or two week period, although the required reading level and length make it challenging for some people to complete.

The Social Rhythm Metric (SRM) quantifies an individual's social zeitgebers (time givers, or circadian rhythm entrainment cues) (Monk et al. 1990). Social life may provide important social cues that entrain circadian rhythms, including sleep habits, eating times, and occupational routines. Disturbance of these social cues could result in dis-entrainment of circadian rhythms, which may increase the risk of developing mood disorders or other mental illnesses. The SRM score is determined from the timing of 15 specific and 2 built-in activities that constitute an individual's social rhythm. If the timing of an activity that occurs at least three times a week is within 45 minutes of the typical time, it is considered part of one's daily routine. The total number of these activities is divided by the total number of activities occurring at least three times a week. The result is the SRM score. A higher SRM score was found to relate to subjective better sleep, higher morning alertness and a deeper nocturnal temperature trough, whereas lower SRM-scores correlated with higher reports of depressive symptoms (Monk et al. 1994).

The Pittsburgh Sleep Quality Index (PSQI) is a self-administered survey which assesses sleep quality and disturbances over a one month time interval (Buysse et al. 1989). Nineteen

individual items generate seven component scores: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The sum of scores for these seven components yields one global score.

The Positive and Negative Syndrome Scale (PANSS) is a 30-question clinician-administered survey that measures symptom severity of schizophrenia and has been widely used in the study of antipsychotic therapy (Kay et al. 1987). Seven questions assess positive symptoms, which refer to an excess or distortion of normal functions, e.g., hallucinations and delusions. Seven questions assess negative symptoms, which represent a decrease or loss of normal function, e.g. blunted affect and social withdrawal. 16 questions assess general psychopathology, e.g. feelings of guilt and poor attention. Each answer is rated 1 to 7 based on the interview as well as reports of family members or healthcare providers. The overall PANSS score thus ranges from 30 to 210. Kay's original publication reported a mean score of 77 for patients with schizophrenia.

The Calgary Depression Scale for Schizophrenia (CDSS) is a nine-item clinician-administered survey, in which each item is on a four point Likert scale (Addington et al. 1990). It was designed to assess depression specifically in psychotic populations, for whom previous depression instruments were not designed. Internal consistency is high, and significant and strong correlations have been found between scores on the CDSS, BDI, and HRSD (Addington et al. 1992; Addington et al. 1993). The CDSS depression score is obtained by adding each of the item scores. A score above 6 has an 82% specificity and 85% sensitivity for detecting a major depressive episode.

The Unified Parkinson's Disease Rating Scale (UPDRS) is a clinician-administered interview and exam that is used to describe the severity of Parkinson's Disease (Fahn et al. 1987). It is made up of the 1) Mentation, Behavior, and Mood, 2) ADL, and 3) Motor sections. Some sections require multiple grades assigned to each extremity. The score ranges from 0 to 199; 0 represents no disability, and 199 represents total disability. Strengths of the UPDRS include its wide utilization, its application across the clinical spectrum of PD, its

nearly comprehensive coverage of motor symptoms, and its clinimetric properties, including reliability and validity. Weaknesses include several ambiguities in the written text, inadequate instructions for raters, some metric flaws, and the absence of screening questions on several important non-motor aspects of PD (Goetz 2003). The motor section of the UPDRS (UPDRS-III) is often used in lieu of the entire UPDRS for patients with PD, and the exam is ideally performed by a movement disorder specialist. In 2007 the Movement Disorder Society (MDS) revised the UPDRS which originally placed nonmotor elements in PD throughout the subscales, with mental features captured in Part I, pain in Part II, and sleep disorders and dysautonomia in Part IV (Goetz et al. 2007). The scale was reorganized so that Part I of the MDS-UPDRS is now titled "Nonmotor Experiences of Daily Living" and encompasses questions requiring medical expertise to answer (cognitive impairment, hallucinations, depressed mood, anxious mood, apathy, and dopamine dysregulation) as well as simpler questions that were considered better suited for a patient or caregiver questionnaire (sleep, staying awake, pain and abnormal sensory sensations, urinary function, constipation, lightheadedness on standing, and fatigue). Part II was retitled to "Motor Experiences of Daily Living", Part III remains "Motor Examination" to be completed by the rater, and Part IV was restricted to "Motor Complications" which include dyskinesias and motor fluctuations. This revised MDS-UPDRS is now commonly used in PD research.

The Hoehn and Yahr (HY) scale was originally designed to be a descriptive, clinician-administered structured interview and staging scale that estimates clinical function in PD, combining functional disability and objective signs of impairment (Hoehn et al. 1967). Strengths of the HY scale include its wide utilization and acceptance. Higher stages correlate with dopaminergic loss as confirmed via neuroimaging studies, and the HY scale has been shown to highly correlate with some standardized scales of motor impairment, disability, and quality of life (Goetz et al. 2004). Weaknesses include the scale's mixing of impairment and disability. Because the HY scale is weighted heavily toward postural instability in determining disease severity, it does not capture impairments or disability from other motor

features of PD, and gives no information on nonmotor problems which are also features of the illness that contribute to decreased quality of life. The UPDRS has largely supplanted the HY scale in clinical and research use.

The Short Form-36 (SF-36) is the most widely used health-related quality-of-life measure in research to date, and can be either self-administered or administered by a trained interviewer over the phone or in person (Ware Jr et al. 1992). The SF-36 yields eight scale scores and two summary scores: a physical component summary (PCS), and mental component summary (MCS). The physical and mental components were designed to be uncorrelated. The eight scale scores represent physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, general mental health, social functioning, energy/fatigue or vitality, and general health perceptions. A higher score represents better health. The PCS and MCS scores are calculated by z-scoring each of the eight scores across the general U.S. population, then multiplying by the corresponding factor scoring coefficient for each scale (Taft et al. 2001).

The Instrumental Activities of Daily Living (IADL) scale assesses independent living skills, identifies how a person is functioning at the present time, and determines improvement or deterioration over time (Lawton et al. 1969). In the original study, the survey was administered by a social worker who gathered information from the subjects, family members, employees, etc. Eight domains of function are measured: ability to use the telephone, shopping, food preparation, housekeeping, laundry, mode of transportation, responsibility for own medications, and ability to handle finances. The IADL Scale is intended to be used among older adults, and may be used in community, clinic, or hospital settings, but is not useful for institutionalized older adults. Although the IADL Scale is easy to administer and focuses on practical functionality related to daily living, it relies on self-report or surrogate report rather than a demonstration of the functional task.

The State Trait Anxiety Inventory (STAI) is a 40-item self-administered survey designed to measure anxiety at two ends of the "affect curve", e.g. feelings of anxiety due to a stressful

state or situation, versus enduring personality traits (Spielberger et al. 1983). Each item has a four-point Likert scale measure. Overall scores thus range from 20 to 80, with higher scores suggesting more severe anxiety.

The Generalized Anxiety Disorder (GAD-7) is a 7-item self-administered survey used to identify GAD (Spitzer et al. 2006). It was constructed from 965 adult primary care patients who completed a questionnaire and telephone interview with a mental health professional within a week, and achieved a sensitivity of 89% and specificity of 82% in assessing generalized anxiety disorder, with good agreement between self-report and interviewer-administered versions of the scale.

The Apathy Evaluation Scale (AES) is used to evaluate apathy – the lack of will to act and the inability to care about the consequences – in a patient based on interview of a person familiar with the patient (Marin 1996). The scale consists of 18 questions that each use a four point Likert scale measure ranging from 0 to 3. Overall scores thus range from 0 to 54; the higher the score the greater the level of apathy.

The Brief Psychiatric Rating Scale (BPRS) is used for measuring general psychiatric symptoms such as depression, anxiety, hallucinations and unusual behavior (Overall et al. 1962). During a structured clinical interview, 18-24 symptoms are scored, and each symptom is rated 1-7 where 1 indicates absence of symptomatology or concern, and 7 indicates extreme severity. The BPRS is one of the oldest and most widely used scales to measure psychotic symptoms.

A.2. Supplemental tables

Table A1: Abbreviations used throughout review

Abbreviation	Definition
AD	Alzheimer's disease
ADHD	Attention deficit hyperactivity disorder
ADL	Activities of daily living
ANS	Autonomic nervous system
AUC	Area under the receiver operating characteristic curve
BD	Bipolar disorder
BLT	Bright light therapy
CDC	Centers for Disease Control and Prevention
DALY	Daily adjusted life year
ECG	Electrocardiogram
EMA	Ecological momentary assessment
GPS	Global positioning system
GSR	Galvanic skin response
HF	High frequency
HRV	Heart rate variability
IS	Interdaily stability
IV	Intradaily variability
L5	Mean activity level during the least active five hours
LF	Low frequency
M10	Mean activity level during the most active ten hours
MCI	Mild cognitive impairment
MDD	Major depressive disorder

Continued on next page

Table A1 – continued from previous page

Acronym	Definition
PPG	Photoplethysmography
PTSD	Post traumatic stress disorder
RA	Relative amplitude
RMSSD	Root mean square of the successive differences
rCBF	Regional cerebral blood flow
SAD	Seasonal affective disorder
SCID	Structured clinical interview for the DSM-IV
SDNN	Standard deviation of average normal-to-normal intervals
SMS	Short message service
VLF	Very low frequency power

Table A2: Aberrations in physiology and behavior associated with neuropsychiatric illness that are detectable by sensors in smartphones and wearables

	Sensor type			
Illness	Accelerometry	HR	GPS	Calls & SMS
Stress & depression	Disruptions in circadian rhythm and sleep	Emotion mediates vagal tone which manifests as altered HRV	Irregular travel routine	Decreased social interactions
Bipolar disorder	Disruptions in circadian rhythm and sleep, locomotor agitation during manic episode	ANS dysfunction via HRV measures	Irregular travel routine	Decreased or increased social interactions
Schizophrenia	Disruptions in circadian rhythm and sleep, locomotor agitation or catatonia, diminished overall activity	ANS dysfunction via HRV measures	Irregular travel routine	Decreased social interactions
PTSD	Inconclusive evidence	ANS dysfunction via HRV measures	Inconclusive evidence	Decreased social interactions
Dementia	Disruptions in circadian rhythm, diminished locomotor activity	Inconclusive evidence	Wandering away from home	Decreased social interaction
Parkinson's disease	Gait impairment, ataxia, dyskinesia	ANS dysfunction via HRV measures	Inconclusive evidence	Voice features can indicate vocal impairment

Table A3: Questionnaires, surveys, and scales

Reference	**Survey (acronym)**	**Indication**
Cohen et al. 1983	Perceived Stress Scale (PSS)	Stress
Hamilton 1960	Hamilton Rating Scale for Depression (HRSD or HAMD)	Depression
Rush et al. 2000	16-item Quick Inventory of Depressive Symptomatology (QIDS-SR16)	Depression
Spitzer et al. 1999	Patient Health Questionnaire (PHQ)	Depressive disorders
Radloff 1977	Center for Epidemiological Studies Depression (CESD)	Depression
Beck et al. 1961	Beck Depression Inventory (BDI)	Depression
Young et al. 1978	Young Mania Rating Scale (YMRS)	Mania
Altman et al. 1997	Altman Self-Rating Mania scale (ASRM)	Mania
Depue et al. 1981	General Behavior Inventory (GBI)	Mania and depression
Monk et al. 1990	Social Rhythm Metric (SRM)	Circadian entrainment
Buysse et al. 1989	Pittsburgh Sleep Quality Index (PSQI)	Sleep
Kay et al. 1987	Positive and Negative Syndrome Scale (PANSS)	Schizophrenia
Addington et al. 1990	Calgary Depression Scale for Schizophrenia (CDSS)	Depression in schizophrenia
Fahn & Elton 1987	Unified Parkinson's Disease Rating Scale (UPDRS)	Parkinson's disease
Hoehn & Yahr 1967	Hoehn and Yahr (HY) scale	Parkinson's disease
Ware & Sherbourne 1992	Short Form-36 (SF-36)	Quality of life
Lawton & Brody 1969	Lawton Instrumental Activities of Daily Living (IADL)	–

Continued on next page

Table A3 – continued from previous page

Reference	Survey (acronym)	Indication
Spielberger et al. 1983	State Trait Anxiety Inventory (STAI)	Anxiety
Spitzer et al. 2006	7-item Generalized Anxiety Disorder (GAD-7) scale	Anxiety
Marin et al. 1996	Apathy Evaluation Scale (AES)	Apathy
Overall & Gorham 1962	Brief Psychiatric Rating Scale (BPRS)	General psychiatric symptoms

Table A4: Studies of smartphones and wearables for monitoring neuropsychiatric illness

Reference	Key aim	Population	Sensors	Design
Abdullah et al. 2016	Estimate social rhythms (assessed via SRM questionnaires) using smartphone data	Seven subjects with BD	Smartphones recorded GPS data, accelerometry, microphone audio, and social communication	Offline retrospective
Aguilera et al. 2015	Assess relationship between daily / weekly mood scores and PHQ-9 scores	33 subjects	Smartphone administered PHQ-9 surveys	Offline retrospective
Albert et al. 2017	Distinguish subjects with PD from controls using accelerometry of hand tremor	Eight subjects with PD and 18 controls	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
AlHanai et al. 2017	Classify subject mood while reading happy or sad stories using wearable data	Ten healthy subjects	Audio was recorded using Apple iPhones. Samsung Simband smartwatches recorded PPG, ECG, accelerometry, skin impedance, galvanic skin response, and skin temperature	Online real-time
Apiquian et al. 2017	Assess motor activity and sleep time before and after antipsychotic treatment	20 subjects with schizophrenia and 20 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Barnett et al. 2018	Predict clinical relapse from behavioral anomalies in two-week window prior to event	17 subjects with schizophrenia	Smartphones recorded mobility, social activity, and questionnaires	Offline prospective
Beiwinkel et al. 2016	Depressive and manic symptoms (assessed via HAMD and YMRS questionnaires administered every three weeks) were classified using smartphone data	13 subjects with BD	Smartphones recorded GPS, accelerometery, and cell tower data; mood states were assessed via a self-reported two-item questionnaire	Offline retrospective
Ben-Zeev et al. 2015	Correlate smartphone features with daily stress ratings, PHQ-9, PSS, and Revised UCLA Loneliness Scale scores	47 healthy subjects	Smartphones recorded GPS, accelerometry, sleep duration, and time proximal to human speech	Offline retrospective
Berle et al. 2010	Assess motor activity and rest-activity characteristics	46 subjects with schizophrenia and 32 controls	Wrist-worn devices recorded actigraphy	Offline retrospective
Bullock et al. 2014	Assess rest-activity metrics in BD patients with low and high trait vulnerability (assessed via the GBI questionnaire)	72 subjects with BD	Wrist-worn devices recorded accelerometry	Offline retrospective
Burns et al. 2011	Correlate EMA survey scores with smartphone features	Eight subjects with MDD	Smartphones recorded GPS, accelerometry, ambient light, and recent calls	Offline retrospective
Canzian et al. 2015	Correlate and predict PHQ score deviations with smartphone features	28 healthy subjects	Smartphones recorded GPS and accelerometry	Offline prospective
Capecci et al. 2016	Identify freezing of gait events using accelerometry	20 subjects with PD	Smartphones recorded accelerometry while subjects walked and were video recorded	Offline retrospective
Cella et al. 2017	Assess autonomic dysfunction in schizophrenia using wearable device data	30 subjects with schizophrenia and 25 controls	Empatica E4 devices recorded skin conductance, HRV, and accelerometry	Offline retrospective
Ellis et al. 2015	Compare outcome measures of gait and gait variability in subjects with PD versus controls	12 subjects with PD and 12 controls	Steps were captured via a smartphone, heel-mounted sensors, and a sensor mat	Offline retrospective

Continued on next page

Table A4 – continued from previous page

Reference	Key aim	Population	Sensors	Design
Kamdar et al. 2016	Estimate variance of emotional state from wearable data via random forest	13 healthy subjects	Samsung Gear S smartwatches recorded accelerometry, ambient light, heart rate; web app administered mood surveys	Offline retrospective
Moore et al. 2012	Forecast mood time series using previous week's self-rated mood data via exponential smoothing and Gaussian process regression	100 subjects with BD	Mood surveys recorded via SMS	Offline prospective
Faedda et al. 2016	Distinguish BD from ADHD using wearables data	48 subjects with BD, 65 subjects with ADHD, and 42 controls	Belt-worn devices recorded accelerometry for five minutes	Offline retrospective
Faurholt-Jepsen et al. 2015	Correlate smartphone data with depressive and manic symptoms via HDRS-17 and YMRS scores assessed monthly	61 subjects with BD	Smartphones recorded speech duration, social activity, and accelerometry	Offline retrospective
Maria et al. 2016	Classify depressive and manic states (via HDRS-17 and YMRS scores) using smartphone data and voice features	28 subjects with BD	Smartphones recorded voice features (pitch, duration, etc.), speech duration, social activity, and accelerometry	Offline retrospective
Fasmer et al. 2015	Fit resting and active periods to power law distributions and assess differences in MDD	47 subjects with MDD and 29 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Griffiths et al. 2012	Assess features of dyskinesia and akinesia from wearable data, and identify improvements in UPDRS scores after medication	34 subjects with PD and 10 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Grünerbl et al. 2015	Depressive and manic symptoms (assessed via HAMD and YMRS questionnaires administered every three weeks) were classified using smartphone data	Ten subjects with BD	Smartphones recorded GPS, accelerometry, number and length of phone calls, and speech and voice features	Offline retrospective
Hauge et al. 2011	Assess motor activity and rest-activity characteristics	24 subjects with schizophrenia, 25 subjects with depression, and 32 controls	Wrist-worn devices recorded actigraphy	Offline retrospective
Kassavetis et al. 2016	Correlate UPDRS scores with smartphone data	14 subjects with PD	Smartphones recorded accelerometry while subjects performed motor tasks	Offline retrospective
Kheirkhahan et al. 2016	Correlate impaired mobility from wearable data	1,135 subjects	Hip-worn devices recorded accelometry	Offline retrospective
Kim et al. 2015	Classify freezing episodes from normal walking using accelerometry	15 subjects with PD	Smartphones recorded accelerometry while subjects walked and were video recorded	Offline retrospective
Kostikis et al. 2014	Correlate accelerometry features with UPDRS hand tremor scores	23 subjects with PD	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
Kostikis et al. 2015	Distinguish subjects with PD from controls using accelerometry of hand tremor	25 subjects and 20 controls	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective

Continued on next page

Table A4 – continued from previous page

Reference	Key aim	Population	Sensors	Design
Krane-gartiser et al. 2014	Assess mean activity, variance, symbolic dynamics, and power spectral features	18 subjects with mania and 12 subjects with BD	Wrist-worn devices recorded accelerometry	Offline retrospective
Kuhlmei et al. 2013	Associate activity with apathy and depression (assessed via AES and BDI questionnaires)	32 subjects with dementia, 21 subjects with MCI, and 23 controls	Wrist-worn devices recorded accelerometry during motor tasks	Offline retrospective
Lee et al. 2015	Compare RR peak detection, HRV measures, and stress detection from wearable versus Holter monitor	17 subjects	Custom ECG patch was developed to record cardiac activity	Offline retrospective
Lee et al. 2016	Correlate UPDRS scores with smartphone data	103 subjects with PD	Smartphones recorded hand dexterity via timed tapping test, rapid alternating movements, tremor tracker via tracing between two parallel lines, and a cognitive interference test	Offline retrospective
Martin et al. 2006	Assess time in bed, sleep consistency, daytime sleeping, and circadian rhythm regularity	28 subjects with schizophrenia and 28 controls	Wrist-worn devices recorded accelerometry and light exposure	Offline retrospective
Nakamura et al. 2007	Fit resting and active periods to power law distributions and assess differences in MDD	14 subjects with MDD and 11 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Nero et al. 2015	Define accelerometer cut points for different walking speeds in adults with PD	30 subjects with PD	Waist-worn devices recorded accelerometry	Offline retrospective
Niwa et al. 2011	Assess if medication status, MMSE scores, activity, and HRV features differed by disease severity (assessed via UPDRS scores) or disease duration	27 subjects with PD and 30 controls	Wrist-worn devices recorded accelerometry and Holter monitors recorded ambulatory ECG	Offline retrospective
O'Brien et al. 2016	Assess relationship between quality of life, ADLs, learning, and depression (assessed via SF-36 and IADLS questionnaires) and smartphone data	29 subjects with MDD and 30 controls	Wrist-worn devices recorded accelerometry. Quality of life, ADLs, learning, and depression were assessed via SF-36 and IADLS questionnaires	Offline retrospective
Osipov et al. 2015	Classify schizophrenic subjects from controls using rest-activity characteristics and HRV features	16 subjects with schizophrenia and 19 controls	Adhesive patches recorded locomotor activity and ECG	Offline retrospective
Palmius et al. 20177	Estimate depressive symptoms (assessed via QIDS-SR16 questionnaires administered weekly) and detect depression using smartphone data	22 subjects with BD and 14 controls	Smartphones recorded GPS data	Offline retrospective
Pan et al. 2015	Correlate accelerometry features with UPDRS scores, and use features to detect hand resting tremor and gait difficulty	40 subjects with PD	Smartphones recorded accelerometry of hand tremor and gait during motor and walking tasks	Offline retrospective
Patel et al. 2009	Estimate UPDRS scores using wearable data	12 subjects with PD	Arm and leg-worn devices recorded accelerometry	Offline retrospective
Place et al. 2017	Estimate depression and PTSD symptoms (assessed via SCID questionnaires) using smartphone data	73 subjects with at least one symptom of PTSD or depression	Smartphones recorded GPS, accelerometry, calls and SMS activity, device use, and voice audio	Offline retrospective

Continued on next page

Table A4 – continued from previous page

Reference	Key aim	Population	Sensors	Design
Reinertsen et al. 2017a	Classify patients with PTSD using time-domain, frequency-domain, and complexity features from RR interval time series	23 subjects with PTSD and 25 controls	A Holter monitor recorded RR intervals for 24 hours	Offline retrospective
Reinertsen et al. 2017b	Classify schizophrenic subjects from controls using rest-activity characteristics and HRV features, and evaluate relationship between number of days of data and classifier accuracy	16 subjects with schizophrenia and 19 controls	Adhesive patches recorded locomotor activity and ECG	Offline retrospective
Roh et al. 2014	Compare RR peak detection, signal-to-noise, and HRV measures from wearable versus Holter monitor	12-41 subjects (varied by test)	Custom ECG patch was developed to record cardiac activity	Offline retrospective
Roy et al. 2011	Classify tremor and dyskinesia from wearable data	11 subjects with PD	Arm and leg-worn devices recorded accelerometry	Offline retrospective
Saeb et al. 2015	Classify low from high PHQ-9 scores using smartphone features	28 healthy subjects	Smartphones recorded GPS and phone usage	Offline retrospective
Saeb et al. 2016	Correlate PHQ-9 scores with smartphone features from weekend vs. weekday data	48 healthy subjects	Smartphones recorded GPS and phone usage	Offline retrospective
Sano et al. 2012	Fit resting and active periods to power law distributions and assess differences in schizophrenia	19 subjects with schizophrenia and 11 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Sano et al. 2013	Distinguish stressed from non-stressed states using wearable data	18 subjects	Wrist-worn devices recorded accelerometry and skin conductance. Smartphones recorded call and SMS activity. Surveys assessed stress, mood, sleep, tiredness, general health, alcohol or caffeine intake, and electronics usage.	Offline retrospective
Sano et al. 2015	Estimate PSQI, PSS, and MCS questionnaire scores from wearable data	66 subjects	Wrist-worn devices recorded accelerometry and skin conductance. Smartphones recorded call and SMS activity. Sleep, stress, and mental health were assessed via PSQI, PSS, and MCS questionnaires respectively	Offline retrospective
Shin et al. 2016	Correlate symptom severity (assessed via the PANSS questionnaire) with activity levels	61 subjects with schizophrenia	Wrist-worn devices recorded accelerometry	Offline retrospective
Stamatakis et al. 2013	Classify UPDRS score categories from wearable data	36 subjects with PD and 10 controls	Finger-worn sensors recorded accelerometry during a tapping test	Offline retrospective
Tung et al. 2014	Compare area, perimeter, and mean distance from home in subjects with AD versus controls using smartphone data	19 subjects with AD and 33 controls	Smartphones recorded GPS	Offline retrospective
Walther et al. 2009b	Assess if motor symptoms (assessed via PANSS questionnaires) correlate with wearables data	55 subjects with schizophrenia	Wrist-worn devices recorded actigraphy	Offline retrospective
Walther et al. 2009a	Assess if activity differs by schizophrenia subtype	60 subjects with schizophrenia	Wrist-worn devices recorded actigraphy	Offline retrospective

Continued on next page

Table A4 – continued from previous page

Reference	Key aim	Population	Sensors	Design
Wang et al. 2014	Correlate smartphone data with PHQ-9, PSS, flourishing scale, and UCLA loneliness scale scores	48 healthy subjects	Smartphones recorded accelerometry, conversations, sleep, and location	Offline retrospective
Wang et al. 2016	Determine associations between EMA survey scores and smartphone data via generalized estimating equations	21 subjects with schizophrenia	Smartphones recorded accelerometry, voice audio, light sensor readings, GPS data, and application usage	Offline retrospective
Weenk et al. 2017	Evaluate association between changes in HRV measures and stress in surgeons	20 subjects	Adhesive patch measured single-lead ECG, respiratory rate, skin temperature, body posture, activity, and steps	Offline retrospective
Wichniak et al. 2011	Measure association between activity levels and mental status (measured via PANSS and CDSS questionnaires)	73 subjects with schizophrenia and 36 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Winkler et al. 2005	Assess if light therapy can improve sleep efficiency and stability in people with seasonal affective disorder (SAD)	17 subjects with SAD and 17 controls	Wrist actigraphy was recorded from which sleep-wake amplitude, phase, and sleep efficiency was estimated	Offline retrospective
Woods et al. 2014	Distinguish PD from essential tremor using accelerometry	14 subjects with PD and 18 subjects with essential tremor	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
Vallance et al. 2011	Assess relationship between depression (assessed via PHQ-9 questionnaires) and activity	2,862 subjects	Wrist-worn devices recorded accelerometry	Offline retrospective

Table A5: Platforms, pilots, and ongoing studies.

Reference or study	Sample size	Methods
Faurholt-Jepsen, M. et al. 2017	400 subjects with BD	Patients will be randomized to either 1) a smartphone-based monitoring system including a feedback loop between patients and clinicians, and cognitive behavioral therapy, or 2) standard treaatment. The outcomes are number and duration of re-admissions, 2) severity of depressive and manic symptoms, and 3) perceived stress, quality of life, symptomatology, etc.
AURORA	5,000 subjects with trauma	Verily, University of North Carolina, and Harvard University are leading a 19-institution five-year endeavor to perform the most comprehensive observational study of trauma to date. Investigators will examine passive data collection methods using smartphone apps, as well as in-person visits, genomic measurements, neurocognitive tests, patient surveys, and medical record reviews. This collaboration presents a unique opportunity to discover new insights that could translate into fundamental advances in our understanding of post-traumatic conditions. See https://www.nimh.nih.gov/news/science-news/2016/nimh-funded-study-to-track-the-effects-of-trauma.shtml.
Healthy Aging Study	100,000 subjects	The overarching goal is to develop a midlife biomarker of Alzheimer's disease, since it is now well established that the disease begins about 2 decades prior to the onset of clinical symptoms. It is critical to develop new ways to detect the disease in the silent asymptomatic phase in order to develop preventative treatments. To accomplish this goal, the Emory Healthy Aging Study first aims to recruit 100,000 individuals to participate in an online study to assess risk factors identified in health questionnaires and by apps to measure cognition. The second aim is to deeply phenotype a subpopulation of about 3000 or more of these subjects every few years to assess a variety of risk factors by profiling genetics, cardiovascular physiology, blood and spinal fluid biomarkers, brain and retinal imaging. Multi-level longitudinal analyses of subjects profiles, including their amyloid status, will facilitate discovery of new biomarkers. See https://healthyaging.emory.edu/about-the-study/.
Batista, E. et al. 2015	16 subjects	Study of AD and MCI. The System for the Private and Autonomous Surveillance based on Information and Communication Technologies (SIMPATIC) project is a smartphone app-based system for monitoring people with MCI. The smartphone app raises alarms under certain conditions, such as an AD patient leaving a defined geographic zone (e.g. home), not moving after a certain amount of time, moving at too high a speed (suggesting they are utilizing transportation), or the phone battery level reaching too low a level.
Faurholt-Jepsen, M. et al. 2013.	78 subjects	Six month study of BD. The "MONARCA" smartphone app administered subjective questionnaires assessing mood, sleep, medicine intake, etc., and monitored speech duration, social activity, and accelerometry.
RADAR-CNS: Remote Assessment of Disease and Relapse - Central Nervous System	Unknown	A collaborative research program exploring the potential of wearable devices to help prevent and treat depression, multiple sclerosis and epilepsy. Jointly led by King's College London and Janssen Pharmaceutica NV, funded by the Innovative Medicines Initiative, and includes 23 organizations from across Europe and the US.
UCLA Depression Grand Challenge	Study aims to enroll 100,000 people	10-year study with aim of identifying will screen for depression, analyze participants' genetics, measure early adversity and life stress and assess symptoms through remote monitoring using cell phones and wearable devices.
mPower: Mobile Parkinson Disease Study	48,000 people downloaded the app; 9,520 people consented to share data	This study will monitor individual's health and symptoms of PD progression like dexterity, balance and gait using questionnaires and sensors via the Parkinson mPower mobile phone application and wearable devices if available.

A.3. Supplemental figures

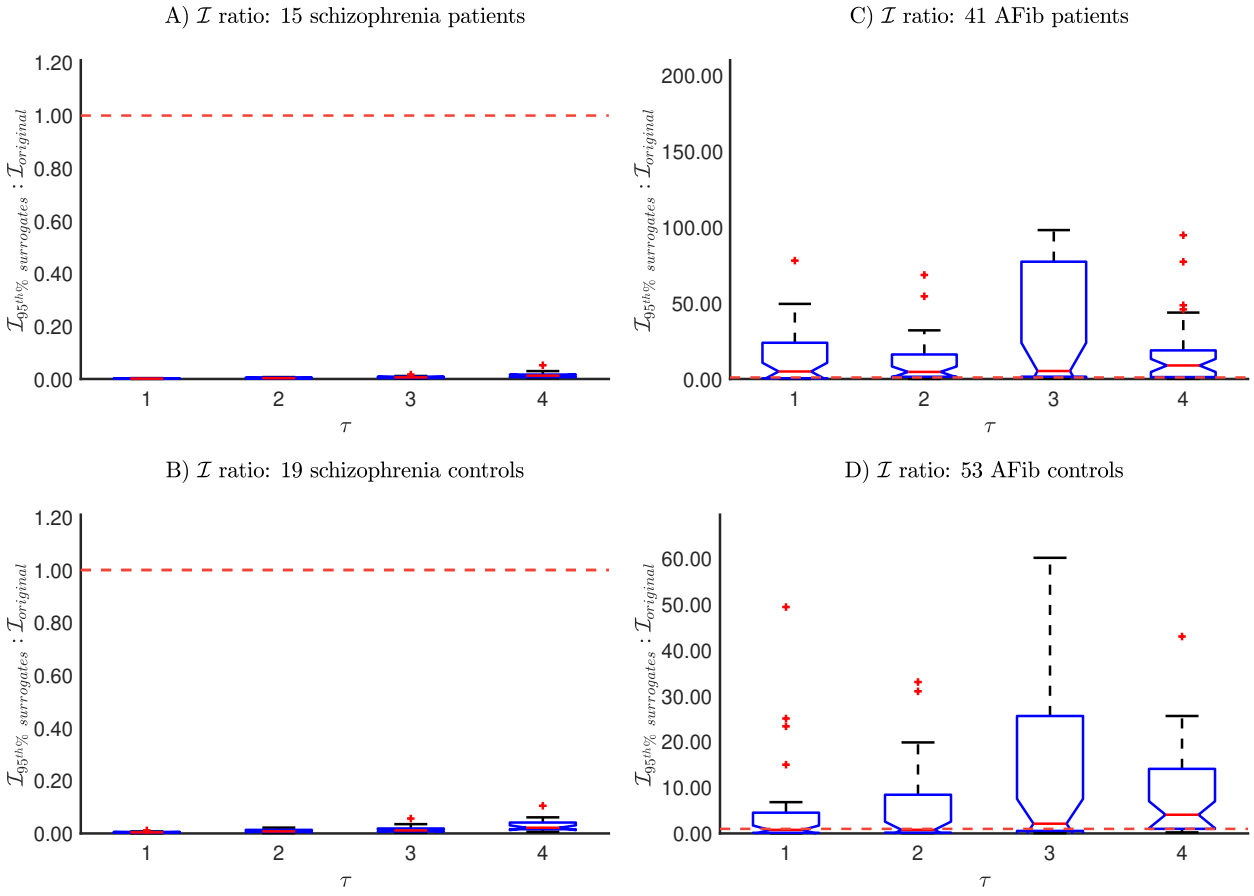


Figure A1: Mutual information ratio; numerator is mutual information between surrogate HR and activity time series generated via random shuffling, and denominator is mutual information between original HR and activity time series. Data is shown via notched box plots; the horizontal red line denotes the median, the notches denote $95th$ percent confidence intervals of the median, the lower and upper blue box denotes the $25th$ and $75th$ percentiles, and the lower and upper whiskers denote the minimum and maximum values, respectively. The horizontal red dashed line indicates unity, i.e. a ratio of 1. A ratio below unity indicates significant mutual information, whereas a ratio equal to or greater than unity indicates the same mutual information is generated from random data. A) Patients in the schizophrenia study have high ratios for all time scales τ, demonstrating significant mutual information compared to random chance, and suggesting coupling between HR and activity. B) Controls in the schizophrenia study have ratios about an order of magnitude lower than controls, although still > 1, suggesting much less coupling between HR and activity in healthy people. C) Patients in the AFib study have low ratios < 1, suggesting observed mutual information is due to random chance. D) Controls in the AFib study also have low ratios.

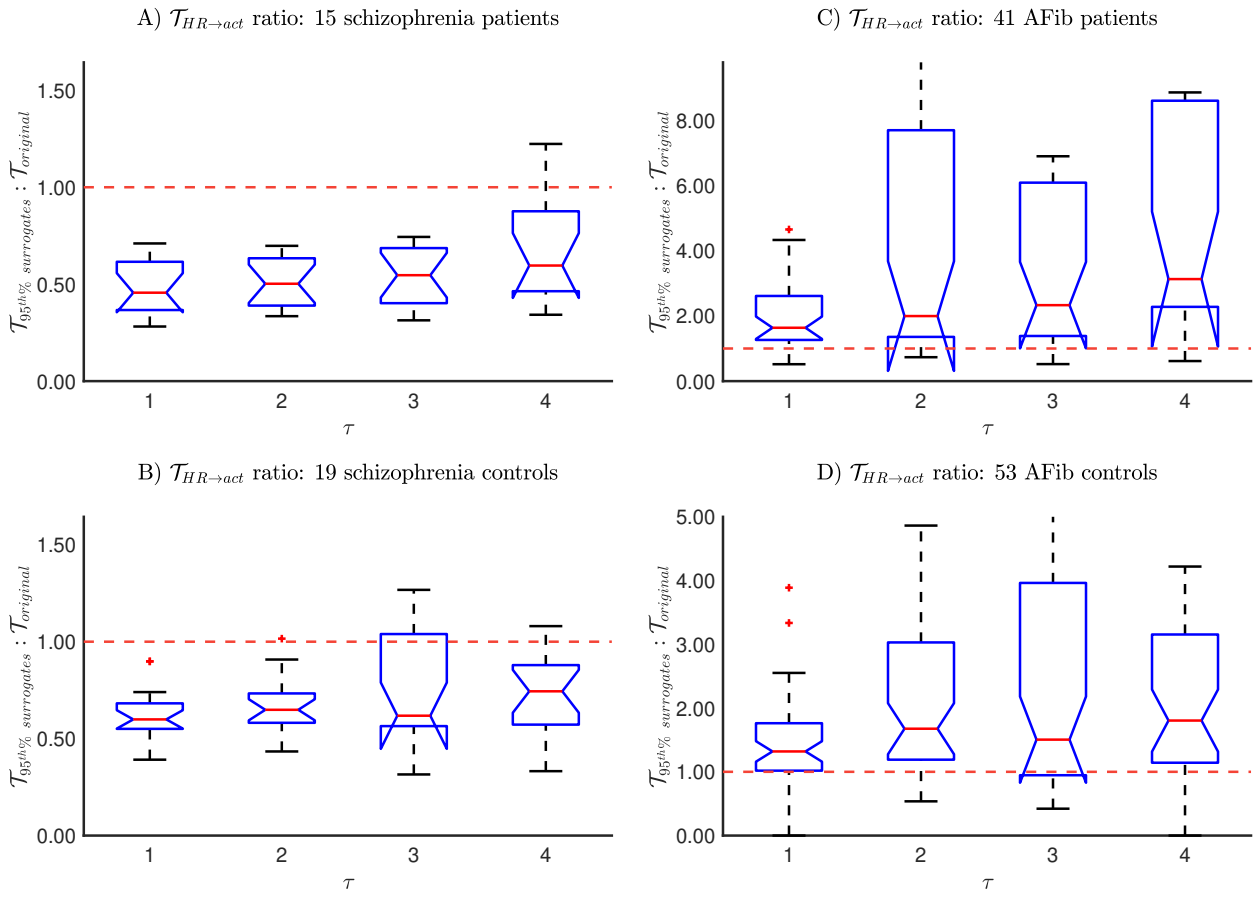


Figure A2: Ratio metric of transfer entropy from HR to activity; numerator is transfer entropy from surrogate HR to activity time series generated via random shuffling, and denominator is transfer entropy from original HR to activity time series. Data is shown via notched box plots; the horizontal red line denotes the median, the notches denote $95th$ percent confidence intervals of the median, the lower and upper blue box denotes the $25th$ and $75th$ percentiles, and the lower and upper whiskers denote the minimum and maximum values, respectively. The horizontal red dashed line indicates unity, i.e. a ratio of 1. A ratio below unity indicates significant directed transfer of information, whereas a ratio equal to or greater than unity indicates the same level of directed information transfer is generated from random data. A) Patients in the schizophrenia study have median transfer entropy ratios below 1 for several time scales, demonstrating significant mutual information compared to random chance, suggesting directed transfer of information from HR to activity. B) Controls in the schizophrenia study have similar transfer entropy ratios. 95% confidence intervals cross 1 for $\tau = 3$ and $\tau = 4$ suggesting less directed information transfer from HR to activity in healthy people at certain time scales. C) Patients in the AFib study have high median transfer entropy ratios, greater than 1 for all time scales, suggesting observed directed information transfer is due to random chance. D) Controls in the AFib study also have median transfer entropy ratios greater than unity.

REFERENCES

Abdullah, S., Matthews, M., and Frank, E. (2016). "Automatic detection of social rhythms in bipolar disorder". *Journal of the American Medical Informatics Association* 23.3, pp. 538–543.

Adams, R. P. and MacKay, D. J. C. (2007). "Bayesian Online Changepoint Detection". *arXiv*, p. 7.

Addington, D., Addington, J., Eleanor, M.-T., et al. (1992). "Reliability and validity of a depression rating scale for schizophrenics". *Schizophrenia Research* 6.3, pp. 201–208.

Addington, D., Addington, J., Maticka-Tyndale, E., et al. (1993). "Assessing depression in schizophrenia: the Calgary Depression Scale". *The British Journal of Psychiatry* 22, pp. 39–44.

Addington, D., Addington, J., and Schissel, B. (1990). "A depression rating scale for schizophrenics". *Schizophrenia research* 3.4, pp. 247–51.

Agelink, M. W., Boz, C., Ullrich, H., et al. (2002). "Relationship between major depression and heart rate variability. Clinical consequences and implications for antidepressive treatment." *Psychiatry Research* 113, pp. 139–149.

Aguilera, A., Schueller, S. M., and Leykin, Y. (2015). "Daily mood ratings via text message as a proxy for clinic based depression assessment". *Journal of Affective Disorders* 175, pp. 471–474.

Airola, A. and Pahikkala, T. (2009). "A comparison of AUC estimators in small-sample studies". *Journal of Machine Learning Research* 8, pp. 3–13.

Akselrod, S., Gordon, D., Ubel, F. F., et al. (1981). "Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control". *Science* 213.4504, pp. 220–22.

Albert, M. V., Toledo, S., Shapiro, M., et al. (2017). "Using Mobile Phones for Activity Recognition in Parkinson's Patients". *Frontiers in Neurology* 3, p. 158.

AlHanai, T., Ghassem, M. M., Waddah, AlHanai, T. W., et al. (2017). "Predicting Latent Narrative Mood using Audio and Physiologic Data". *AAAI*.

Allen, J. (2017). "Photoplethysmography and its application in clinical physiological measurement". *Physiological Measurement*.

Altman, E. G., Hedeker, D., Peterson, J. L., et al. (1997). "The Altman Self-Rating Mania Scale". *Biological psychiatry* 42.10, pp. 948–955.

Alvares, G. A., Quintana, D. S., Hickie, I. B., et al. (2016). "Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis". *Journal of Psychiatry & Neuroscience* 41.2, pp. 89–104.

Alzheimer's Association (2016). *2016 Alzheimer's disease facts and figures*. Tech. rep. 4, pp. 459–509.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. 5th ed. Arlington, VA: American Psychiatric Publishing.

Andrews, S., Ellis, D. A., Shaw, H., et al. (2015). "Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use". *PLoS ONE* 10.10, e0139004.

Apiquian, R., Fresan, A., Jairo, M.-D., et al. (2017). "Variations of rest-activity rhythm and sleep-wake in schizophrenic patients versus healthy subjects: An actigraphic comparative study". *Biological Rhythm Research* 39.1, pp. 69–78.

Appel, M. L., Berger, R. D., Saul, J. P., et al. (1989). "Beat to beat variability in cardiovascular variables: Noise or music?" *Journal of the American College of Cardiology* 14.5, pp. 1139–1148.

Arlot, S. and Celisse, A. (2010). "A survey of cross-validation procedures for model selection". *Statistics Surveys* 4, pp. 40–79.

Asch, D. A., Muller, R. W., and Volpp, K. G. (2012). "Automated Hovering in Health Care - Watching Over the 5000 Hours". *New England Journal of Medicine* 367.1, pp. 1–3.

Aung, M., Matthews, M., and Choudhury, T. (2017). "Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies". *Depression and Anxiety* 34.7, pp. 603–609.

Bagby, R., Ryder, A., Schuller, D., et al. (2017). "The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight?" *American Journal of Psychiatry* 161.12, pp. 2163–2177.

Bär, K. J., Letzsch, A., Jochum, T., et al. (2005). "Loss of efferent vagal activity in acute schizophrenia". *Journal of Psychiatric Research* 39.5, pp. 519–27.

Bär, K. J., Wernich, K., Boettger, S., et al. (2008). "Relationship between cardiovagal modulation and psychotic state in patients with paranoid schizophrenia". *Psychiatry Research* 157, pp. 255–257.

Bär, K.-J., Boettger, M. K., Koschke, M., et al. (2017). "Non-linear complexity measures of heart rate variability in acute schizophrenia". *Clinical Neurophysiology* 118.9, pp. 2009–2015.

Barbieri, R. and Brown, E. N. (2008). "Application of dynamic point process models to cardiovascular control". *BioSystems* 93.1-2, pp. 120–125.

Barnett, I., Torous, J., Staples, P., et al. (2018). "Relapse prediction in schizophrenia through digital phenotyping: a pilot study". *Neuropsychopharmacology* January, p. 1.

Barrett, P. M., Steinhubl, S. R., Muse, E. D., et al. (2017). "Digitising the mind". *Lancet* 389.10082, p. 1877.

Bauer, A., Kantelhardt, J. W., Barthel, P., et al. (2006a). "Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study". *Lancet* 367.9523, pp. 1674–81.

Bauer, A., Kantelhardt, J. W., Bunde, A., et al. (2006b). "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data". *Physica A: Statistical Mechanics and its Applications* 364, pp. 423–434.

Beauchaine, T. (2001). "Vagal tone, development, and Gray's motivational theory: toward an integrated model of autonomic nervous system functioning in psychopathology". *Development and Psychopathology* 13, pp. 183–214.

Beck, A. T., Steer, R. A., and Antonio, B. G. K. (1996). "Beck depression inventory-II". *San Antonio* 78.2, pp. 490–498.

Beck, A. T., Ward, C. H., Mendelson, M., et al. (1961). "An inventory for measuring depression". *Archives of General Psychiatry* 4.6, pp. 561–571.

Behar, J., Roebuck, A., Domingos, J. S., et al. (2013). "A review of current sleep screening applications for smartphones". *Physiological Measurement* 34.7, R29–R46.

Beiwinkel, T., Kindermann, S., Maier, A., et al. (2016). "Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study". *JMIR Mental Health* 3.1, e2.

Ben-Zeev, D., Scherer, E. A., Wang, R., et al. (2015). "Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health." *Psychiatric Rehabilitation Journal* 3.3, pp. 218–226.

Berle, J. O., Hauge, E. R., Oedegaard, K. J., et al. (2010). "Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression". *BMC Research Notes* 3.1, p. 149.

Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., et al. (2000). "Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability". *Journal of the American College of Cardiology* 35.6, pp. 1462–1469.

Bigger, J. T., Albrecht, P., Steinman, R. C., et al. (1989). "Comparison of Time- and Frequency Domain-Based Measures of Cardiac Parasympathetic Activity in Holter Record-

ings After Myocardial Infarction". *The American Journal of Cardiology* 64.8, pp. 536–538.

Bigger, J. T., Fleiss, J. L., Steinman, R. C., et al. (1992). "Frequency domain measures of heart period variability and mortality after myocardial infarction." *Circulation* 85.1, pp. 164–71.

Billman, G. E. and Dujardin, J. P. (1990). "Dynamic changes in cardiac vagal tone as measured by time-series analysis". *The American journal of physiology* 258.3, pp. 896–902.

Billman, G. E. (2013). "The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance". *Frontiers in Physiology* 4 FEB.February, pp. 1–5.

Birkhofer, A., Geissendoerfer, J., Alger, P., et al. (2013). "The deceleration capacity - a new measure of heart rate variability evaluated in patients with schizophrenia and antipsychotic treatment". *European Psychiatry* 28.2, pp. 81–86.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Clarendon Press, p. 482.

Bloomfield, D. M., Kaufman, E. S., Bigger, J. T., et al. (1997). "Passive head-up tilt and actively standing up produce similar overall changes in autonomic balance". *American Heart Journal* 134.2, pp. 316–320.

Bot, B. M., Suver, C., Neto, E. C., et al. (2016). "The mPower study, Parkinson disease mobile data collected using ResearchKit". *Scientific Data* 3, p. 160011.

Breiman, L. (1996). "Bagging predictors". *Machine Learning* 24.2, pp. 123–140.

Buckley, P. F., Miller, B. J., Lehrer, D. S., et al. (2009). "Psychiatric comorbidities and schizophrenia". *Schizophrenia Bulletin* 35.2, pp. 383–402.

Bullock, B. and Murray, G. (2014). "Reduced amplitude of the 24 hour activity rhythm: a biomarker of vulnerability to bipolar disorder?" *Clinical Psychological Science* 2.1, pp. 86–96.

Burns, M. N., Begale, M., Duffecy, J., et al. (2011). "Harnessing context sensing to develop a mobile intervention for depression." *Journal of Medical Internet Research* 13.3, e55.

Buysse, D., Reynolds, C. F., Monk, T. H., et al. (1989). "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research". *Psychiatry Research*.

Byerly, M. J., Nakonezny, P. A., and Lescouflair, E. (2007). "Antipsychotic medication adherence in schizophrenia". *Psychiatric Clinics of North America* 30.3, pp. 437–452.

Cakmak, A., Reinertsen, E., Nemati, S., et al. (2018). "Benchmarking changepoint detection algorithms on cardiac time series". *Submitted*.

Campana, L. M., Owens, R. L., Clifford, G. D., et al. (2010). "Phase-rectified signal averaging as a sensitive index of autonomic changes with aging". *Journal of Applied Physiology* 108.6, pp. 1668–1673.

Campanharo, A. S. L. O., Sirer, M. I., De Malmgren, R. D., et al. (2011). "Duality between time series and networks". *PLoS ONE* 6.8, pp. 1–12.

Canzian, L. and Musolesi, M. (2015). "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis". *UbiComp 2015*, pp. 1293–1304.

Capecci, M., Pepa, L., Verdini, F., et al. (2016). "A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease". *Gait & Posture* 50, pp. 28–33.

Carpenter, L. L., Tyrka, A. R., J, McDougle, C., et al. (2004). "Cerebrospinal fluid corticotropin-releasing factor and perceived early-life stress in depressed patients and healthy control subjects". *Neuropsychopharmacology* 29.4, p. 777.

Cella, M., Okruszek, Ł., Lawrence, M., et al. (2017). "Using wearable technology to detect the autonomic signature of illness severity in schizophrenia". *Schizophrenia Research*.

Cemer, I. (2011). *Noise Measurement*.

Chang, C.-C. and Lin, C.-J. (2011). "LIBSVM: a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology* 2.3, p. 27.

Chang, J. S., Yoo, C. S., Yi, S. H., et al. (2009). "Differential pattern of heart rate variability in patients with schizophrenia". *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 33.6, pp. 991–995.

Chen, W., Wang, Z., Xie, H., et al. (2007). "Characterization of surface EMG signal based on fuzzy entropy". *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 15.2, pp. 266–272.

Chow, S.-C. (2014). "Adaptive Clinical Trial Design". *Annual Review of Medicine* 65.1, pp. 405–415.

Clifford, G. D. (2002). "Signal processing methods for heart rate variability". PhD thesis. University of Oxford.

Clifford, G. D. (2006). "ECG statistics, noise, artifacts, and missing data". *Advanced Methods and Tools for ECG Data Analysis*. Artech House, pp. 55–99.

Clifford, G. D. (2016). "The use of sustainable and scalable health care technologies in developing countries". *Innovation and Entrepreneurship in Health* 3, pp. 35–46.

Clifford, G. D., Behar, J., Li, Q., et al. (2012a). "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms". *Physiological Measurement* 33.9, pp. 1419–1433.

Clifford, G. D. and Clifton, D. A. (2012b). "Wireless Technology in Disease Management and Medicine". *Annual Review of Medicine* 63.1, pp. 479–492.

Clifford, G. D., Lopez, D., Li, Q., et al. (2011). "Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments". *Computing in Cardiology* 1419, pp. 285–288.

Clifford, G. D. and Tarassenko, L. (2005). "Quantifying errors in spectral estimates of HRV due to beat replacement and resampling". *IEEE Transactions on Biomedical Engineering* 52.4, pp. 630–638.

Clifford, G. D., Tarassenko, L., Gari D Clifford, et al. (2004). "Segmenting cardiac-related data using sleep stages increases separation between normal subjects and apnœic patients". *Physiological Measurement* 25.6, pp. 27–35.

Cohen, H., Benjamin, J., Geva, A. B., et al. (2000). "Autonomic dysregulation in panic disorder and in post-traumatic stress disorder: application of power spectrum analysis of heart rate variability at rest and in response to recollection of trauma or panic attacks". *Psychiatry Research* 96.1, pp. 1–13.

Cohen, S., Kamarck, T., Mermelstein, R., et al. (1983). "A global measure of perceived stress". *Journal of Health and Social Behavior* 24.4, pp. 385–396.

Collins, P. Y., Patel, V., Joestl, S. S., et al. (2011). "Grand challenges in global mental health". *Nature* 475.7354, pp. 27–30.

Copeland, L. A., Zeber, J. E., Salloum, I. M., et al. (2017). "Treatment Adherence and Illness Insight in Veterans With Bipolar Disorder". *The Journal of Nervous and Mental Disease* 196.1, p. 16.

Costa, M., Goldberger, A., and Peng, C.-K. (2002). "Multiscale entropy analysis of complex physiologic time series". *Physical Review Letters*.

Depue, R. A., Slater, J. F., Heidi, W.-K., et al. (1981). "A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: A conceptual framework and five validation studies." *Journal of Abnormal Psychology* 90.5, p. 381.

Dohrenwend, B. P., Turner, J. B., Turse, N. a., et al. (2006). "The psychological risks of Vietnam for U.S. veterans: a revisit with new data and methods". *Science* 313.5789, pp. 979–82.

Draghici, A. E. and Taylor, J. A. (2016). "The Physiological Basis and Measurement of Heart Rate Variability in Humans". *Journal of Physiological Anthropology* 35.1, pp. 22–29.

Duncan, T. E. (1970). "On the calculation of mutual information". *SIAM Journal on Applied Mathematics* 19.1, pp. 215–220.

Eadicicco, L. (2016). "Americans Check Their Phones 8 Billion Times a Day". *Time*.

Ebner-Priemer, U. W., Kuo, J., Welch, S., et al. (2006). "A valence-dependent group-specific recall bias of retrospective self-reports: A study of borderline personality disorder in everyday life". *The Journal of Nervous and Mental Disease* 194.10, pp. 774–779.

Ehrenreich, B., Righter, B., Rocke, D., et al. (2017). "Are Mobile Phones and Handheld Computers Being Used to Enhance Delivery of Psychiatric Treatment?: A Systematic Review". *The Journal of Nervous and Mental Disease* 199.11, p. 886.

Ellis, R. J., Ng, Y., Zhu, S., et al. (2015). "A validated smartphone-based assessment of gait and gait variability in Parkinson's disease". *PLoS ONE* 10.10, e0141694.

Emory University (2016). *Emory Healthy Aging Study*.

Espel, E. S., Blackburn, E. H., Lin, J., et al. (2004). "Accelerated telomere shortening in response to life stress". *Proceedings of the National Academy of Sciences* 101.49.

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". *Nature* 542.7639, pp. 115–118.

Faedda, G. L., Ohashi, K., Hernandez, M., et al. (2016). "Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls". *Journal of Child Psychology and Psychiatry* 57.6, pp. 706–716.

Fahn, S. and Elton, R. L. (1987). "Unified Parkinsons Disease Rating Scale". *Recent developments in Parkinson's disease*. Vol. 2. Macmillan Healthcare Information, pp. 153–163.

Fasmer, E. E., Fasmer, O. B., Berle, J. O., et al. (2018). "Graph theory applied to the analysis of motor activity in patients with schizophrenia and depression". *PLoS ONE* 13.4, e0194791.

Fasmer, O. B., Hauge, E., Berle, J., et al. (2015). "Distribution of Active and Resting Periods in the Motor Activity of Patients with Depression and Schizophrenia". *Psychiatry Investigation* 13.1, pp. 112–120.

Faurholt-Jepsen, M., Vinberg, M., Frost, M., et al. (2015). "Smartphone data as an electronic biomarker of illness activity in bipolar disorder". *Bipolar Disorders* 17.7, pp. 715–728.

Femminella, G. D., Rengo, G., Kimici, K., et al. (2014). "Autonomic dysfunction in Alzheimer's disease: tools for assessment and review of the literature". *Journal of Alzheimer's Disease* 42, pp. 369–377.

Firth, J. A. J. J. A., Cotter, J., Torous, J., et al. (2016). "Mobile Phone Ownership and Endorsement of "mHealth" Among People With Psychosis: A Meta-analysis of Cross-sectional Studies". *Schizophrenia Bulletin* 42.2, pp. 448–455.

Free, C., Phillips, G., Galli, L., et al. (2013). "The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review." *PLOS Medicine* 10.1, e1001362.

Freeman, D., Sheaves, B., Goodwin, G. M., et al. (2017). "The effects of improving sleep on mental health (OASIS): a randomised controlled trial with mediation analysis." *Lancet Psychiatry* 4.10, pp. 749–758.

Furlan, R., Guzzetti, S., Crivellaro, W., et al. (1990). "Continuous 24-hour assessment of the neural regulation of systemic arterial pressure and RR variabilities in ambulant subjects". *Circulation* 81.2, pp. 537–547.

Galatzer-Levy, I. R., Karstoft, K. I., Statnikov, A., et al. (2014). "Quantitative forecasting of PTSD from early trauma responses: A Machine Learning application". *Journal of Psychiatric Research* 59, pp. 68–76.

Gandhi, M. and Wang, T. (2014). *The Future of Biosensing Wearables*. Tech. rep.

Gederi, E., Nemati, S., Edwards, B. A., et al. (2014). "Model-based estimation of loop gain using spontaneous breathing: A validation study". *Respiratory Physiology and Neurobiology* 201, pp. 84–92.

Germain, A. (2013). "Sleep Disturbances as the Hallmark of PTSD: where are we now?" *American Journal of Psychiatry* April, pp. 372–382.

Germain, A., Hall, M., Krakow, B., et al. (2005). "A brief Sleep Scale for Posttraumatic Stress Disorder: Pittsburgh Sleep Quality Index Addendum for PTSD". *Journal of Anxiety Disorders* 19.2, pp. 233–244.

Ghassemi, M., Lehman, L. H. L.-w., Snoek, J., et al. (2014). "Global optimization approaches for parameter tuning in biomedical signal processing: a focus of multi-scale entropy". *Computers in Cardiology*, 41:993–996.

Goetz, C. G. (2003). "The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations". *Movement Disorders* 18.7, pp. 738–750.

Goetz, C. G., Fahn, S., Martinez-Martin, P., et al. (2007). "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Process, format, and clinimetric testing plan". *Movement Disorders* 22.1, pp. 41–47.

Goetz, C. G., Poewe, W., Rascol, O., et al. (2004). "Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations". *Movement Disord* 19.9, pp. 1020–1028.

Gracious, B., Youngstrom, E. R. A., Findling, R., et al. (2002). "Discriminative Validity of a Parent Version of the Young Mania Rating Scale". *Journal of the American Academy of Child & Adolescent Psychiatry* 41.11, pp. 1350–1359.

Griffiths, R. I., Kotschet, K., Arfon, S., et al. (2012). "Automated assessment of bradykinesia and dyskinesia in Parkinson's disease". *Journal of Parkinson's Disease* 2.1, pp. 47–55.

Grippo, A. J. and Johnson, A. K. (2009). "Stress, depression and cardiovascular dysregulation: a review of neurobiological mechanisms and the integration of research from preclinical disease models". *Stress* 12.1, pp. 1–21.

Grünerbl, A., Muaremi, A., Osmani, V., et al. (2015). "Smartphone-based recognition of states and state changes in bipolar disorder patients". *IEEE Journal of Biomedical and Health Informatics* 19.1, pp. 140–148.

Gulshan, V., Peng, L., Coram, M., et al. (2016). "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *JAMA* 304.6, pp. 649–656.

Haaksma, J., Dijk, W., Brouwer, J., et al. (1998). "The influence of recording length on time and frequency domain analysis of heart rate variability". *Computers in Cardiology*, pp. 377–380.

Haensel, A., Mills, P. J., Nelesen, R. A., et al. (2008). "The relationship between heart rate variability and inflammatory markers in cardiovascular diseases". *Psychoneuroendocrinology* 33.10, pp. 1305–1312.

Hall, J. (2010). *Guyton and Hall Textbook of Medical Physiology*. 12th ed. Elsevier.

Hamilton, M. (1960). "A rating scale for depression". *Journal of Neurology, Neurosurgery, and Psychiatry* 23, pp. 56–62.

Hattori, S., Kishida, I., Suda, A., et al. (2017). "Effects of four atypical antipsychotics on autonomic nervous system activity in schizophrenia". *Schizophrenia Research*.

Hauge, E. R., Berle, J. O., Oedegaard, K. J., et al. (2011). "Nonlinear analysis of motor activity shows differences between schizophrenia and depression: a study using Fourier analysis and sample entropy". *PLOS ONE* 6.1, e16291.

Hausdorff, J. M., Peng, C. K., Ladin, Z., et al. (1995). "Is walking a random walk? Evidence for long-range correlations in stride interval of human gait". *Journal of Applied Physiology* 78.1, pp. 349–358.

Henry, B. L., Minassian, A., Paulus, M. P., et al. (2010). "Heart rate variability in bipolar mania and schizophrenia". *Journal of Psychiatric Research* 44.3, pp. 168–76.

Hoehn, M. M. and Yahr, M. D. (1967). "Parkinsonism: onset, progression and mortality". *Neurology* 17.5, pp. 427–442.

Holtzman, C., Trotman, H., Goulding, S., et al. (2013). "Stress and neurodevelopmental processes in the emergence of psychosis". *Neuroscience* 249, pp. 172–191.

Honey, G. D., Pomarol-Clotet, E., Corlett, P. R., et al. (2005). "Functional dysconnectivity in schizophrenia associated with attentional modulation of motor function". *Brain* 128.11, pp. 2597–2611.

Hor, K. and Taylor, M. (2010). "Suicide and schizophrenia: a systematic review of rates and risk factors." *Journal of psychopharmacology* 20.4, pp. 81–90.

Huang, W.-l., Chang, L.-r., Kuo, T. B., et al. (2013). "Impact of Antipsychotics and Anticholinergics on Autonomic Modulation in Patients With Schizophrenia". *Journal of Clinical Psychopharmacology* 33.2, pp. 170–177.

Hudson, J. E. (2006). "Signal Processing Using Mutual Information". *IEEE Signal Processing Magazine* 23.6, pp. 50–54.

Insel, T. R. (2017). "Digital Phenotyping: Technology for a New Science of Behavior". *JAMA*.

Ivanov, P. C., Amaral, L. a. N., Goldberger, A. L., et al. (1999). "Multifractality in human heartbeat dynamics". *Nature* 399.6735, pp. 461–465.

Johnson, A. E., Ghassemi, M. M., Nemati, S., et al. (2016). "Machine Learning and Decision Support in Critical Care". *Proceedings of the IEEE* 104.2.

Jongsma, H. E., Gayer-Anderson, C., Lasalvia, A., et al. (2018). "Treated incidence of psychotic disorders in the multinational EU-GEI study". *JAMA Psychiatry* 75.1, pp. 36–46.

Julien, C. (2006). "The enigma of Mayer waves: Facts and models". *Cardiovascular Research* 70.1, pp. 12–21.

Kahn, R. S., Sommer, I. E., Murray, R. M., et al. (2015). "Schizophrenia". *Nature Reviews Disease Primers* 1, nrdp201567.

Kamdar, M. R. and Wu, M. J. (2016). "Prism: a data-driven platform for monitoring mental health". *Pacific Symposium on Biocomputing* 21, pp. 333–344.

Kantelhardt, J. W., Bauer, A., Schumann, A. Y., et al. (2007). "Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk". *Chaos* 17.1, pp. 1–9.

Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis*. Cambridge University Press.

Karam, E. G., Friedman, M. J., Hill, E. D., et al. (2014). "Cumulative Traumas And Risk Thresholds: 12-Month PTSD In The World Mental Health (WMH) Surveys". *Depression and Anxiety* 31.2, pp. 130–142.

Karemaker, J. (2017). "An introduction into autonomic nervous function". *Physiological Measurement* 38, aa6782.

Karow, A., Pajonk, F.-G., Reimer, J., et al. (2008). "The dilemma of insight into illness in schizophrenia: self- and expert-rated insight and quality of life". *European Archives of Psychiatry and Clinical Neuroscience* 258.3, pp. 152–159.

Karstoft, K.-I., Galatzer-Levy, I. R., Statnikov, A., et al. (2015). "Bridging a translational gap: using machine learning to improve the prediction of PTSD". *BMC Psychiatry* 15, p. 30.

Kassavetis, P., Saifee, T. A., Roussos, G., et al. (2016). "Developing a Tool for Remote Digital Assessment of Parkinson's Disease". *Movement Disorders Clinical Practice* 3.1, pp. 59–64.

Katona, P. G., Poitras, J. W., Barnett, G. O., et al. (1970). "Cardiac vagal efferent activity and heart period in the carotid sinus reflex". *American Journal of Physiology-Legacy Content* 218.4, pp. 1030–1037.

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". *Schizophrenia Bulletin* 13.2, pp. 261–276.

Kazdin, A. E. and Rabbitt, S. M. (2013). "Novel models for delivering mental health services and reducing the burdens of mental illness". *Clinical Psychological Science* 1.2, pp. 170–191.

Kemp, A. H., Quintana, D. S., Gray, M. a., et al. (2010). "Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis". *Biological Psychiatry* 67.11, pp. 1067–1074.

Kessler, R. C., Rose, S., Koenen, K. C., et al. (2014). "How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys." *World Psychiatry* 13.3, pp. 265–74.

Kheirkhahan, M., Catrine, T.-L., Axtell, R., et al. (2016). "Actigraphy features for predicting mobility disability in older adults". *Physiological Measurement* 37.10, pp. 1813–1833.

Khoury, N. M., Marvar, P. J., Gillespie, C. F., et al. (2012). "The renin-angiotensin pathway in posttraumatic stress disorder: angiotensin-converting enzyme inhibitors and angiotensin receptor blockers are associated with fewer traumatic stress symptoms." *The Journal of Clinical Psychiatry* 73.6, pp. 849–55.

Kim, H., Lee, H., Lee, W., et al. (2015). "Unconstrained detection of freezing of gait in Parkinson's disease patients using smartphone". *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2015, pp. 3751–3754.

Kirkpatrick, B. and Miller, B. J. (2013). "Inflammation and schizophrenia". *Schizophrenia Bulletin* 39.6, pp. 1174–1179.

Kleiger, R. E., Stein, P. K., Bosner, M. S., et al. (1992). "Time domain measurements of heart rate variability". *Cardiology Clinics* 10.3, pp. 487–98.

Kleiger, R. E., Stein, P. K., and Bigger, J. T. (2005). "Heart Rate Variability: Measurement and Clinical Utility". *Annals of Noninvasive Electrocardiology* 10.1, pp. 88–101.

Knoepfli-Lenzin, C., Sennhauser, C., Toigo, M., et al. (2010). "Effects of a 12-week intervention period with football and running for habitually active men with mild hypertension". *Scandinavian Journal of Medicine & Science in Sports* 20, pp. 72–79.

Kobayashi, I., Lavela, J., and Mellman, T. A. (2014). "Nocturnal Autonomic Balance and Sleep in PTSD and Resilience". *Journal of Traumatic Stress* 27.6, pp. 712–716.

Kok, B. C., Herrell, R. K., Thomas, J. L., et al. (2012). "Posttraumatic stress disorder associated with combat service in Iraq or Afghanistan: reconciling prevalence differences between studies." *The Journal of Nervous and Mental Disease* 200.5, pp. 444–50.

Kostikis, N., Hristu-Varsakelis, D., Arnaoutoglou, M., et al. (2015). "A smartphone-based tool for assessing parkinsonian hand tremor". *IEEE Journal of Biomedical and Health Informatics* 19.6, pp. 1835–1842.

Kostikis, N., Hristu-Varsakelis, M., Arnaoutoglou, M., et al. (2014). "Smartphone-based Evaluation of Parkinsonian hand tremor: Quantitative Measurements vs Clinical Assessment Scores". *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 906–909.

Krane-gartiser, K., Henriksen, T., Morken, G., et al. (2014). "Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder". *PLoS ONE* 9.2, e89574.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). "The PHQ-9: validity of a brief depression severity measure". *Journal of General Internal Medicine*.

Kuhlmei, A., Walther, B., Becker, T., et al. (2013). "Actigraphic daytime activity is reduced in patients with cognitive impairment and apathy". *European Psychiatry* 28.2, pp. 94–97.

Kuhn, E., Greene, C., Hoffman, J., et al. (2014). "Preliminary Evaluation of PTSD Coach, a Smartphone App for Post-Traumatic Stress Symptoms". *Military Medicine* 179.1, pp. 12–18.

Lan, B. L., Yeoh, E. V., and Ng, J. A. (2010). "Distribution of detrended stock market data". *Fluctuation and Noise Letters* 09.03, pp. 245–257.

Lanata, A., Valenza, G., Nardelli, M., et al. (2015). "Complexity index from a personalized wearable monitoring system for assessing remission in mental health". *IEEE Journal of Biomedical and Health Informatics* 19.1, pp. 132–139.

Lawton, M. P. and Brody, E. M. (1969). "Assessment of older people: self-maintaining and instrumental activities of daily living". *The Gerontologist* 9.3, pp. 179–186.

Lee, H. M., Chen, C. M., Chen, J. M., et al. (2001). "An efficient fuzzy classifier with feature selection based on fuzzy entropy". *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 31.3, pp. 426–432.

Lee, J., Nemati, S., Silva, I., et al. (2012). "Transfer Entropy Estimation and Directional Coupling Change Detection in Biomedical Time Series". *Biomedical Engineering Online* 11.1, p. 19.

Lee, W. K., Yoon, H., and Park, K. S. (2015). "Smart ECG Monitoring Patch with Built-in R-Peak Detection for Long-Term HRV Analysis". *Annals of Biomedical Engineering* 44.7, pp. 2292–2301.

Lee, W., Evans, A., and Williams, D. R. (2016). "Validation of a Smartphone Application Measuring Motor Function in Parkinson's Disease". *Journal of Parkinson's Disease* 6.2, pp. 371–382.

Lewinsohn, P. M., Seeley, J. R., Roberts, R. E., et al. (1997). "Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults". *Psychology and Aging* 12.2, p. 277.

Li, P., Liu, C., Li, L., et al. (2013). "Multiscale multivariate fuzzy entropy analysis". *Acta Physica Sinica* 62.12, p. 120512.

Li, Q., Mark, R. G., and Clifford, G. D. (2008). "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter". *Physiological Measurement* 29.1, pp. 15–32.

Liddell, B. J., Kemp, A. H., Steel, Z., et al. (2016). "Heart rate variability and the relationship between trauma exposure age, and psychopathology in a post-conflict setting". *BMC Psychiatry* 16, p. 133.

Lip, G. Y. H., Fauchier, L., Freedman, S. B., et al. (2016). "Atrial fibrillation". *Nature Reviews Disease Primers* 2, p. 16016.

Liu, C., Li, K., Zhao, L., et al. (2013). "Analysis of heart rate variability using fuzzy measure entropy". *Computers in Biology and Medicine* 43.2, pp. 100–108.

Liu, C., Oster, J., Reinertsen, E., et al. (2018). "Comparison of entropy approaches for atrial fibrillation discrimination". *Physiological Measurement* 39.7, p. 074002.

Lo, A., Chernoff, H., Zheng, T., et al. (2015). "Why significant variables aren't automatically good predictors". *Proceedings of the National Academy of Sciences* 112.45, pp. 13892–13897.

Maes, M., Van Bockstaele, D. R., Gastel, A. V., et al. (1999). "The effects of psychological stress on leukocyte subset distribution in humans: evidence of immune activation". *Biological Psychiatry* 39.1, pp. 1–9.

Maetzler, W., Domingos, J., Srulijes, K., et al. (2013). "Quantitative wearable sensors for objective assessment of Parkinson's disease". *Movement Disorders* 28.12, pp. 1628–1637.

Malarkey, W. B., Pearl, D. K., Demers, L. M., et al. (1995). "Influence of academic stress and season on 24-hour mean concentrations of ACTH, cortisol, and beta-endorphin". *Psychoneuroendocrinology* 20.5, pp. 499–508.

Mancia, G. (2012). "Short- and long-term blood pressure variability: present and future". *Hypertension* 60.2, pp. 512–517.

Manuca, R. and Savit, R. (1996). "Stationarity and nonstationarity in time series analysis". *Physica D: Nonlinear Phenomena* 99.2-3, pp. 134–161.

Maria, F.-J., Busk, J., Frost, M., et al. (2016). "Voice analysis as an objective state marker in bipolar disorder". *Translational Psychiatry* 6.7, e856.

Marin, R. S. (1996). "Apathy: Concept, Syndrome, Neural Mechanisms, and Treatment". *Seminars in Clininical Neuropsychiatry* 1.4, pp. 304–314.

Marmar, C. R., Schlenger, W., Henn-Haase, C., et al. (2015). "Course of posttraumatic stress disorder 40 years after the vietnam war: Findings from the National Vietnam Veterans Longitudinal Study". *JAMA Psychiatry* 10016.9, pp. 875–881.

Martel, T. F. van de (2008). "Faking it: social desirability response bias in self-report research". *Australian Journal of Advanced Nursing* 25.4, pp. 40–48.

Martin, A., Rief, W., Klaiberg, A., et al. (2006). "Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population". *General Hospital Psychiatry* 28.1, pp. 71–77.

Martin, J. L., Jeste, D. V., and Sonia, A.-I. (2005). "Older schizophrenia patients have more disrupted sleep and circadian rhythms than age-matched comparison subjects". *Journal of Psychiatric Research* 39.3, pp. 251–259.

Massachusetts Institute of Technology (2011). *Matlab Tools for Network Analysis*.

McConnell, M. V., Shcherbina, A., Pavlovic, A., et al. (2017). "Feasibility of obtaining measures of lifestyle from a smartphone app: the MyHeart Counts Cardiovascular Health Study". *JAMA Cardiology* 2.1, p. 67.

McCraty, R., Atkinson, M., Tomasino, D., et al. (2001). "Analysis of twenty-four hour heart rate variability in patients with panic disorder". *Biological Psychology* 56.2, pp. 131–150.

Mcguire, T. G. and Miranda, J. (2008). "New Evidence Regarding Racial And Ethnic Disparities In Mental Health: Policy Implications". *Health Affairs* 27.2, pp. 393–403.

Mietus, J. E., Peng, C.-K., Henry, I., et al. (2002). "The pNNx files: re-examining a widely used heart rate variability measure". *Heart* 88.4, pp. 378–80.

Millar, A., Espie, C. A., and Scott, J. (2004). "The sleep of remitted bipolar outpatients: A controlled naturalistic study using actigraphy". *Journal of Affective Disorders* 80.2-3, pp. 145–153.

Minassian, A., Geyer, M. A., Baker, D. G., et al. (2014). "Heart rate variability characteristics in a large group of active-duty Marines and relationship to posttraumatic stress". *Psychosomatic Medicine* 76.4, pp. 292–301.

Mohr, D. C., Zhang, M., and Schueller, S. M. (2016). "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning". *Annual Review of Clinical Psychology* 13.1.

Monasterio, V., Burgess, F., and Clifford, G. D. (2012). "Robust classification of neonatal apnoea-related desaturations". *Physiological Measurement* 33.9, pp. 1503–1516.

Mondelli, V., Dazzan, P., Hepgul, N., et al. (2010). "Abnormal cortisol levels during the day and cortisol awakening response in first-episode psychosis: The role of stress and of antipsychotic treatment". *Schizophrenia Research* 116.2-3, pp. 234–242.

Monk, T. H., Petrie, S. R., Hayes, A. J., et al. (1994). "Regularity of daily life in relation to personality, age, gender, sleep quality and circadian rhythms". *Journal of Sleep Research* 3.4, pp. 196–205.

Monk, T. K., Flaherty, J. F., Frank, E., et al. (1990). "The Social Rhythm Metric: An instrument to quantify the daily rhythms of life". *Journal of Nervous and Mental Disease* 178.2.

Montano, N., Porta, A., Cogliati, C., et al. (2009). "Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior". *Neuroscience and Biobehavioral Reviews* 33.2, pp. 71–80.

Montaquila, J. M., Trachik, B. J., and Bedwell, J. S. (2015). "Heart rate variability and vagal tone in schizophrenia: A review". *Journal of Psychiatric Research* 69.September, pp. 57–66.

Moore, P. J., Little, M. A., E, McSharry, P., et al. (2012). "Forecasting depression in bipolar disorder". *IEEE Transactions on Biomedical Engineering* 59.10, pp. 2801–2807.

Nakamura, T., Kiyono, K., Yoshiuchi, K., et al. (2007). "Universal scaling law in human behavioral organization". *Physical Review Letters* 99.13, p. 138103.

National Institute of Mental Health (2016). *NIMH-Funded Study to Track the Effects of Trauma*.

Nelson, C. R. and Kang, H. (1981). "Spurious periodicity in inappropriately detrended time series". *Econometrica* 49.3, p. 741.

Nemati, S., Edwards, B. A., Lee, J., et al. (2013). "Respiration and heart rate complexity: Effects of age and gender assessed by band-limited transfer entropy". *Respiratory Physiology & Neurobiology* 189.1, pp. 159–163.

Nero, H. akan, Wallén, M., Franzén, E., et al. (2015). "Accelerometer cut points for physical activity assessment of older adults with Parkinson's disease". *PLoS ONE* 10.9, e0135899.

Newcomer, J. W. and Hennekens, C. H. (2007). "Severe mental illness and risk of cardiovascular disease". *JAMA* 298.15, pp. 1794–1796.

NICE guideline (CG178) (2014). *Psychosis and schizophrenia in adults: prevention and management*. Tech. rep.

Niwa, F., Kuriyama, N., Nakagawa, M., et al. (2011). "Circadian rhythm of rest activity and autonomic nervous system activity at different stages in Parkinson's disease". *Autonomic Neuroscience* 165.2, pp. 195–200.

Noah, B., Keller, M. S., Mosadeghi, S., et al. (2017). "Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials". *npj Digital Medicine* 1.1, p. 2.

Novartis (2018). *Novartis expands alliance with Science 37 to advance virtual clinical trials program*.

Obermeyer, Z. and Emanuel, E. J. (2016). "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine". *New England Journal of Medicine* 375.13, pp. 1216–1219.

O'Brien, J., Gallagher, P., Stow, D., et al. (2016). "A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression". *Psychological Medicine* 47.1, pp. 93–102.

O'Regan, C., Kenny, R. A., Cronin, H., et al. (2015). "Antidepressants strongly influence the relationship between depression and heart rate variability: Findings from The Irish Longitudinal Study on Ageing (TILDA)". *Psychological Medicine* 45.3, pp. 623–636.

Osipov, M., Behzadi, Y., Kane, J. M., et al. (2015). "Objective identification and analysis of physiological and behavioral signs of schizophrenia". *Journal of Mental Health* 24.5, pp. 276–282.

Overall, J. E. and Gorham, D. R. (1962). "The Brief Psychiatric Rating Scale". *Psychological Reports* 10.3, pp. 799–812.

Pagani, M., Furlan, R., Dell'Orto, S., et al. (1985). "Simultaneous analysis of beat by beat systemic arterial pressure and heart rate variabilities in ambulatory patients." *Journal of Hypertension* 3.3, S83–5.

Palmius, N., Osipov, M., Bilderbeck, A. C., et al. (2014). "A multi-sensor monitoring system for objective mental health management in resource constrained environments". *Appropriate Healthcare Technologies for Low Resource Settings*, pp. 1–4.

Palmius, N., Tsanas, A., Saunders, K. E. A., et al. (20177). "Detecting Bipolar Depression From Geographic Location Data". *IEEE Transactions on Biomedical Engineering* 64.8, pp. 1761–1771.

Pan, D., Dhall, R., Lieberman, A., et al. (2015). "A mobile cloud-based Parkinson's disease assessment system for home-based monitoring." *JMIR mHealth and uHealth* 3.1, e29.

Pan, Q., Zhou, G., Wang, R., et al. (2016). "Do the deceleration/acceleration capacities of heart rate reflect cardiac sympathetic or vagal activity? A model study". *Medical and Biological Engineering and Computing*, pp. 1–13.

Parati, G., Ochoa, J. E., Lombardi, C., et al. (2015). "Blood Pressure Variability: Assessment, Predictive Value, and Potential as a Therapeutic Target". *Current Hypertension Reports* 17.4, pp. 1–18.

Patel, M. S., Asch, D. A., and Volpp, K. G. (2015). "Wearable Devices as Facilitators, Not Drivers,of Health Behavior Change". *Journal of the American Medical Association* 19104.5, pp. 459–460.

Patel, S., Lorincz, K., Hughes, R., et al. (2009). "Monitoring Motor Fluctuations in Patients with Parkinson's Disease Using Wearable Sensors". *IEEE Transactions on Information Technology in Biomedicine* 13.6, pp. 864–873.

Pedro, B.-G., Ivanov, P., Amaral, L. A., et al. (2001). "Scale Invariance in the Nonstationarity of Human Heart Rate". *Physical Review Letters* 87.16, p. 168105.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., et al. (2008). "Evaluating the added predictive ability of a newmarker: From area under the ROC curve to reclassification and beyond". *Statistics in medicine* 27, pp. 157–172.

Peng, H., Long, F., and Ding, C. (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8, pp. 1226–1238.

Place, S., Danielle, B.-H., Rubin, C., et al. (2017). "Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders". *Journal of Medical Internet Research* 19.3, pp. 1–9.

Poushter, J. (2016). *Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies*. Tech. rep.

Rachow, T., Berger, S., Boettger, M. K., et al. (2011). "Nonlinear relationship between electrodermal activity and heart rate variability in patients with acute schizophrenia". *Psychophysiology* 48, pp. 1323–1332.

Radloff, L. S. (1977). "The CES-D scale: A Self-Report Depression Scale for Research in the General Population". *Applied Psychological Measurement* 1.3, pp. 385–401.

Raffalovich, L. E. (1994). "Detrending time series". *Sociological Methods & Research* 22.4, pp. 492–519.

Rechlin, T., Claus, D., Weis, M., et al. (1994). "Heart rate variability in schizophrenic patients and changes of autonomic heart rate parameters during treatment with clozapine". *Biological Psychiatry* 35.11, pp. 888–892.

Reinertsen, E., Nemati, S., Vest, A. N., et al. (2017a). "Heart rate-based window segmentation improves accuracy of classifying posttraumatic stress disorder using heart rate variability measures". *Physiological Measurement* 38.6, pp. 1061–1076.

Reinertsen, E., Osipov, M., Liu, C., et al. (2017b). "Continuous assessment of schizophrenia using heart rate and accelerometer data". *Physiological Measurement* 38.7, pp. 1456–1471.

Reinertsen, E., Shashikumar, S. P., Shah, A. J., et al. (2018). "Multiscale network dynamics between heart rate and locomotor activity are altered in schizophrenia". *Submitted*.

Resnick, H. S., Kilpatrick, D. G., Dansky, B. S., et al. (1993). "Prevalence of civilian trauma and posttraumatic stress disorder in a representative national sample of women". *Journal of Consulting and Clinical Psychology* 61.6, pp. 984–991.

Richman, J. S. and Moorman, J. R. (2000). "Physiological time-series analysis using approximate entropy and sample entropy". *American Journal of Physiology - Heart and Circulatory Physiology* 278.6, H2039–49.

Rodgers, M. M., Pai, V. M., and Conroy, R. S. (2015). "Recent Advances in Wearable Sensors for Health Monitoring". *IEEE Sensors Journal* 15.6, pp. 3119–3126.

Roh, T., Hong, S., and Yoo, H.-J. (2014). "Wearable depression monitoring system with heart-rate variability". *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 562–565.

Roy, S. H., Cole, B. T., Gilmore, D. L., et al. (2011). "Resolving Signal Complexities for Ambulatory Monitoring of Motor Function in Parkinson's Disease". *2011 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2011, pp. 4836–4839.

Rush, A. J., Carmody, T. J., and Reimitz, P. P.-E. P. (2000). "The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and Self-Report (IDS-SR) ratings of depressive symptoms". *International Journal of Methods in Psychiatric Research* 9.2, pp. 45–59.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., et al. (2003). "The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression". *Biological Psychiatry* 54.5, pp. 573–583.

Sadeh, A. and Acebo, C. (2002). "The role of actigraphy in sleep medicine". *Sleep Medicine Reviews* 6.2, pp. 113–124.

Saeb, S., Lattie, E. G., Schueller, S. M., et al. (2016). "The relationship between mobile phone location sensor data and depressive symptom severity". *PeerJ* 4, e2537.

Saeb, S., Zhang, M., Karr, C. J., et al. (2015). "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study". *Journal of Medical Internet Research* 17.7, e175.

Saha, S., Chant, D., Welham, J., et al. (2005). "A systematic review of the prevalence of schizophrenia". *PLoS Medicine* 2.5, p. 141.

Sano, A., Phillips, A. J., Yu, A. Z., et al. (2015). "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones". *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks*, pp. 1–6.

Sano, A. and Picard, R. W. (2013). "Stress recognition using wearable sensors and mobile phones". *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction Stress*.

Sano, W., Nakamura, T., Yoshiuchi, K., et al. (2012). "Enhanced persistency of resting and active periods of locomotor activity in schizophrenia". *PLoS ONE* 7.8, e43539.

Savage, N. (2015). "Mobile data: Made to measure". *Nature* 527.7576, S12–S13.

Saxena, S., Thornicroft, G., Knapp, M., et al. (2007). "Resources for mental health: scarcity, inequity, and inefficiency". *Lancet* 370.9590, pp. 878–889.

Sayers, J. (2001). "The world health report 2001 - Mental health: new understanding, new hope." *Bulletin of the World Health Organization* 79.11, p. 1085.

Schreiber, T. (2000). "Measuring Information Transfer". *Physical Review Letters* 85.2, p. 461.

Schueller, S. M., Begale, M., Penedo, F. J., et al. (2014). "Purple: A Modular System for Developing and Deploying Behavioral Intervention Technologies". *Journal of Medical Internet Research* 16.7, e181.

Seyfert-Margolis, V. (2018). "The evidence gap". *Nature Biotechnology* 36.3, pp. 228–232.

Shah, A. J., Lampert, R., Goldberg, J., et al. (2013). "Posttraumatic stress disorder and impaired autonomic modulation in male twins". *Biological Psychiatry* 73.11, pp. 1103–1110.

Shahriari, B., Swersky, K., Wang, Z., et al. (2016). "Taking the human out of the loop: a review of Bayesian optimization". *Proceedings of the IEEE*.

Shashikumar, S. P., Li, Q., Clifford, G. D., et al. (2017a). "Multiscale network representation of physiological time series for early prediction of sepsis". *Physiological Measurement*.

Shashikumar, S. P., Shah, A. J., Li, Q., et al. (2017b). "A Deep Learning Approach to Monitoring and Detecting Atrial Fibrillation using Wearable Technology". *IEEE EMBS International Conference on Biomedical & Health Informatics*, pp. 141–144.

Shin, S., Yeom, C.-W., Shin, C., et al. (2016). "Activity monitoring using a mHealth device and correlations with psychopathology in patients with chronic schizophrenia". *Psychiatry Research* 246, pp. 712–718.

Singh, D., Vinod, K., Saxena, S. C., et al. (2004). "Effects of RR segment duration on HRV spectrum estimation". *Physiological Measurement* 25.3, pp. 721–735.

Sokolove, P. G. and Bushell, W. N. (1978). "The chi square periodogram: its utility for analysis of circadian rhythms". *Journal of Theoretical Biology* 72.1, pp. 131–160.

Solhan, M. B., Trull, T. J., Jahng, S., et al. (2009). "Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall". *Psychological Assessment* 21.3, pp. 425–436.

Sowden, G. and Huffman, J. C. (2009). "The impact of mental illness on cardiac outcomes: a review for the cardiologist". *International Journal of Cardiology*.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., et al. (1983). *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., et al. (1999). "Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study". *JAMA* 282.18, pp. 1737–1744.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., et al. (2006). "A Brief Measure for Assessing Generalized Anxiety Disorder". *Archives of Internal Medicine* 166.10, p. 1092.

Stamatakis, J., Ambroise, J., Crémers, J., et al. (2013). "Finger Tapping Clinimetric Score Prediction in Parkinson's Disease Using Low-Cost Accelerometers". *Computational Intelligence and Neuroscience*, pp. 1–13.

Staples, P., Torous, J., Barnett, I., et al. (2017). "A comparison of passive and active estimates of sleep in a cohort with schizophrenia". *npj Schizophrenia* 3.1, p. 37.

Statista (2017). *Wearable device sales revenue worldwide from 2016 to 2022 (in billion U.S. dollars).*

Stein, P. K., Bosner, M. S., Kleiger, R. E., et al. (1994). "Heart rate variability: A measure of cardiac autonomic tone". *American Heart Journal* 127.5, pp. 1376–1381.

Steinhubl, S. R., Muse, E. D., and Topol, E. J. (2015). "The emerging field of mobile health". *Science Translational Medicine* 7.283, 283rv3.

Stilo, S. A. and Murray, R. M. (2010). "The epidemiology of schizophrenia: replacing dogma with knowledge". *Dialogues in clinical neuroscience* 12.3, pp. 305–15.

Su, S., Lampert, R., Lee, F., et al. (2010). "Common genes contribute to depressive symptoms and heart rate variability: The twins heart study". *Twin Research and Human Genetics* 13.1, pp. 1–9.

Sundin, J., Herrell, R. K., Hoge, C. W., et al. (2014). "Mental health outcomes in US and UK military personnel returning from Iraq". *British Journal of Psychiatry* 204.3, pp. 200–207.

Taft, C., Karlsson, J., and Sullivan, M. (2001). "Do SF-36 summary component scores accurately summarize subscale scores?" *Quality of Life Research* 10.5, pp. 395–404.

Takens, F. (1981). "Detecting strange attractors in turbulence". *Lecture notes in mathematics* 898.1, pp. 366–381.

Tan, G., Dao, T. K., Farmer, L., et al. (2011). "Heart rate variability (HRV) and posttraumatic stress disorder (PTSD): A pilot study". *Applied Psychophysiology Biofeedback* 36.1, pp. 27–35.

Tan, G., Fink, B., Dao, T. K., et al. (2009). "Associations among pain, PTSD, mTBI, and heart rate variability in veterans of operation enduring and Iraqi Freedom: A pilot study". *Pain Medicine* 10.7, pp. 1237–1245.

Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use". *Circulation* 17.5, pp. 354–381.

Teicher, M. H. (1995). "Actigraphy and motion analysis: new tools for psychiatry". *Harvard Review of Psychiatry* 3.1, pp. 18–35.

Thayer, J. F., Åhs, F., Fredrikson, M., et al. (2012). "A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health". *Neuroscience and Biobehavioral Reviews* 36.2, pp. 747–756.

Thayer, J. F., Yamamoto, S. S., and Brosschot, J. F. (2010). "The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors". *International Journal of Cardiology* 141.2, pp. 122–131.

Tison, G. H., Sanchez, J. M., Ballinger, B., et al. (2018). "Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch". *JAMA Cardiology* 0124, pp. 1–8.

Torous, J., Chan, S., Tan, S., et al. (2014). "Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics". *JMIR Mental Health* 1.1, e5.

Torous, J., Kiang, M. V., Lorme, J., et al. (2016). "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research." *JMIR Mental Health* 3.2, e16.

Trajković, G., Starčević, V., Latas, M., et al. (2011). "Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years". *Psychiatry Research* 189.1, pp. 1–9.

Tsanas, A., Saunders, K. E. A. K., Bilderbeck, A. C. A. A. C., et al. (2016). "Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder". *Journal of Affective Disorders* 205, pp. 225–233.

Tung, J. Y., Rose, R. V., Gammada, E., et al. (2014). "Measuring life space in older adults with mild-to-moderate Alzheimer's disease using mobile phone GPS". *Gerontology* 60.2, pp. 154–162.

Vallance, J. K., Winkler, E., Gardiner, P. A., et al. (2011). "Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005–2006)". *Preventative Medicine* 53.4-5, pp. 284–288.

Van Someren, E. J. W., Swaab, D. F., Colenda, C. C., et al. (1999). "Bright Light Therapy: Improved Sensitivity to Its Effects on Rest-Activity Rhythms in Alzheimer Patients by

Application of Nonparametric Methods". *The Journal of Biological and Medical Rhythm Research* 16.4, pp. 505–518.

Vancampfort, D., Firth, J., Schuch, F., et al. (2017). "Sedentary behavior and physical activity levels in people with schizophrenia, bipolar disorder and major depressive disorder: a global systematic review and meta-analysis". *World Psychiatry* 16.3, pp. 308–315.

Vanoli, E., Adamson, P. B., Pinna, G. D., et al. (1995). "Heart rate variability during specific sleep stages". *Circulation* 91, pp. 1918–1922.

Vest, A. N., Li, Q., Liu, C., et al. (2017). "Benchmarking heart rate variability toolboxes". *Journal of Electrocardiology* 50.6, pp. 744–747.

Vigo, D., Thornicroft, G., and Atun, R. (2016). "Estimating the true global burden of mental illness". *The Lancet Psychiatry* 3.2, pp. 171–178.

Viola, A. U., Simon, C., Ehrhart, J., et al. (2002). "Sleep processes exert a predominant influence on the 24-h profile of heart rate variability". *Journal of Biological Rhythms* 17.6, pp. 539–547.

Walker, E. F., Trotman, H. D., Pearce, B. D., et al. (2013). "Cortisol levels and risk for psychosis: Initial findings from the North American Prodrome Longitudinal Study". *Biological Psychiatry* 74.6, pp. 410–417.

Walther, S., Horn, H., Razavi, N., et al. (2009a). "Quantitative motor activity differentiates schizophrenia subtypes". *Neuropsychobiology* 60.2, pp. 80–86.

Walther, S., Koschorke, P., Horn, H., et al. (2009b). "Objectively measured motor activity in schizophrenia challenges the validity of expert ratings". *Psychiatry Research* 169.3, pp. 187–190.

Wang, R., Aung, M. S. H., Abdullah, S., et al. (2016). "CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia". *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

Wang, R., Chen, F., Chen, Z., et al. (2014). "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones". *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14.

Ware Jr, J. E. and Sherbourne, C. D. (1992). "The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection". *Medical Care* 131.6, pp. 473–483.

Weenk, M., Alken, A., Engelen, L., et al. (2017). "Stress measurement in surgeons and residents using a smart patch". *American Journal of Surgery*.

Wescott, T. (2010). *Sampling: What Nyquist Didn't Say, and What to Do About It*. Tech. rep.

Whiteford, H. A., Degenhardt, L., Rehm, J., et al. (2013). "Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010". *Lancet* 382.9904, pp. 1575–1586.

Wichniak, A., Skowerska, A., Jolanta, C.-W., et al. (2011). "Actigraphic monitoring of activity and rest in schizophrenic patients treated with olanzapine or risperidone". *Journal of Psychiatric Research* 45.10, pp. 1381–1386.

Winkler, D., Pjrek, E., Nicole, P.-R., et al. (2005). "Actigraphy in Patients with Seasonal Affective Disorder and Healthy Control Subjects Treated with Light Therapy". *Biological Psychiatry* 58.4, pp. 331–336.

Winter, D. A., Quanbury, A. O., and Reimer, G. D. (1972). "Analysis of instantaneous energy of normal gait". *Journal of Biomechanics*.

Witting, W., Kwa, I. H., Eikelenboom, P., et al. (1990). "Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease". *Biological Psychiatry* 27.6, pp. 563–572.

Woodcock, J., Whyte, J., and Henderson, M. B. (2017). *2015-2016 Global Participation in Clinical Trials Report*. Tech. rep.

Woods, A. M., Nowostawski, M., Franz, E. A., et al. (2014). "Parkinson's disease and essential tremor classification on mobile device". *Pervasive and Mobile Computing* 13, pp. 1–12.

Woodward, S. H., Arsenault, N. J., Voelker, K., et al. (2009). "Autonomic Activation During Sleep in Posttraumatic Stress Disorder and Panic: A Mattress Actigraphic Study". *Biological Psychiatry* 66.1, pp. 41–46.

Wu, E. Q., Birnbaum, H. G., Shi, L., et al. (2005). "The economic burden of schizophrenia in the United States in 2002". *Journal of Clinical Psychiatry* 66.9, pp. 1122–1129.

Wu, Z., Huang, N. E., Long, S. R., et al. (2007). "On the trend, detrending, and variability of nonlinear and nonstationary time series". *Proceedings of the National Academy of Sciences of the United States of America* 104.38, pp. 14889–94.

Wulff, K., Dijk, D.-J., Middleton, B., et al. (2012). "Sleep and circadian rhythm disruption in schizophrenia". *The British Journal of Psychiatry* 200.4, pp. 308–316.

Xiong, W., Faes, L., and Ivanov, P. C. (2017). "Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations". *Physical Review E* 95.6, pp. 1–37.

Yehuda, R., Hoge, C. W., McFarlane, A. C., et al. (2015). "Post-traumatic stress disorder". *Nature Reviews Disease Primers* October, p. 15057.

Yesavage, J. A., Kinoshita, L. M., Noda, A., et al. (2014). "Longitudinal assessment of sleep disordered breathing in Vietnam veterans with post-traumatic stress disorder". *Nature and Science of Sleep* 6, pp. 123–127.

Young, R. C., Biggs, J. T., Ziegler, V. E., et al. (1978). "A rating scale for mania: reliability, validity and sensitivity". *The British Journal of Psychiatry* 133, pp. 429–435.

Youngstrom, E. A., Frazier, T. W., Demeter, C., et al. (2008). "Developing a ten item mania scale from the Parent General Behavior Inventory for children and adolescents". *Journal of Clinical Psychiatry* 69.5, pp. 831–839.

Zhao, L., Wei, S., Zhang, C., et al. (2015). "Determination of sample entropy and fuzzy measure entropy parameters for distinguishing congestive heart failure from normal sinus rhythm subjects". *Entropy*, p. 6288.

Zyl, L. T. van, Hasegawa, T., and Nagata, K. (2008). "Effects of antidepressant treatment on heart rate variability in major depression: A quantitative review". *BioPsychoSocial Medicine* 2.12, pp. 1–10.